VALIDATION OF MOUNTAIN PRECIPITATION FORECASTS FROM THE NCAR CONVECTION-PERMITTING ENSEMBLE AND OPERATIONAL FORECAST SYSTEMS OVER THE WESTERN UNITED STATES

by

Thomas Michael Gowan

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Atmospheric Sciences

The University of Utah

December 2017

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of **Thomas Michael Gowan**

has been approved by the following supervisory committee members:

W. James Steenburgh	, Chair	**7/26/2017** Date Approved
John D. Horel	, Member	**7/26/2017** Date Approved
Courtenay Strong	, Member	**7/26/2017** Date Approved

and by **Kevin Perry** , Chair/Dean of

the Department/College/School of **Atmospheric Sciences**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Convection-permitting ensembles (CPEs) can capture the large spatial variability and quantify the inherent uncertainty of precipitation forecasts in areas of complex terrain; however, such systems remain largely untested over the western U.S. In this study, we assess the capabilities of deterministic and probabilistic cool-season (October-March) quantitative precipitation forecasts (QPFs) produced by the high-resolution (3-km horizontal grid spacing), 10-member NCAR Ensemble using observations collected by Snow Telemetry (SNOTEL) stations at mountain locations across the western U.S and precipitation analyses from the Parameter-elevation Relationships on Independent Slopes Model (PRISM). We also examine the performance of operational forecast systems run by the National Centers for Environmental Prediction (NCEP) including the HRRR, NAM 3-km CONUS nest, GFS, and SREF. Overall, we find that higher resolution models, such as the HRRR, NAM-3km CONUS nest, and an individual member of the NCAR Ensemble, are more skillful than coarser models, especially over the interior ranges of the western U.S. This is likely because the high-resolution models better resolve topography, especially the narrow interior ranges, and thus better simulate orographic precipitation. Although probabilistic forecasts from the SREF are often more skillful than those generated by the NCAR Ensemble, the NCAR Ensemble generally outperforms its individual dynamical cores. While the NCAR Ensemble is shown to suffer from a spread deficiency, the SREF's multidynamical core configuration allows it to generate ample

spread. These results should help guide future short-range model development and inform forecasters about the capabilities and limitations of several widely used deterministic and probabilistic modeling systems over the western U.S.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank several people and research groups for their help and guidance. First and foremost, I would like to thank my advisor, Jim Steenburgh. Not only has he been an exceptional advisor and mentor to me throughout the thesis process, but also he has provided me with many great opportunities, including presenting at conferences and visiting and collaborating with groups at NCAR and NCEP. Academically, I have progressed much more than I expected to over the past couple of years and that is largely due to him. I would also like to thank Craig Schwartz, Glen Romine, Ryan Sobash, and Kate Fossell from the NCAR Ensemble team for access to NCAR Ensemble forecast data and answers to all questions regarding the NCAR Ensemble and precipitation verification. During my visit to NCAR, Craig's expertise in ensemble modeling and forecast verification proved to be invaluable. His guidance and ideas dramatically improved the quality of my thesis. Additionally, I would like to thank Eric Rogers and the NAM team for pre-operational NAM-3km forecast data and Yan Lou and the GEFS and GFS team for 13-km GFS data. Lastly, I would also like to thank my committee, John Horel and Court Strong, and the entire Mountain Meteorology group for their help and support during the thesis process.

This thesis is based on research supported by the NOAA/National Weather Service CSTAR Program through Grants NA13NWS4680003 and NA17NWS4680001. Any

opinions, findings, and conclusions or recommendations expressed herein are those of the author and do not necessarily reflect those of the NOAA/National Weather Service.

CHAPTER 1

INTRODUCTION

Recent increases in computational capabilities have allowed for the development of ensemble numerical weather prediction (NWP) modeling systems with horizontal grid spacings ≤ 4 km, such that cumulous parameterizations can me removed (Kain et al. 2008). Commonly referred to as "convection-permitting" ensembles (CPEs), these modeling systems offer significant promise for improving quantitative precipitation forecasts (QPFs) and probabilistic QPFs (PQPFs) over the western U.S. At present, deterministic convection-permitting models (CPMs) run operationally by the National Centers for Environmental Prediction (NCEP), such as the High Resolution Rapid Refresh (HRRR) and North American Mesoscale Forecast System 3-km CONUS nest (hereafter NAM-3km), provide high-resolution numerical guidance but no information concerning forecast uncertainty, except in a time-lagged sense (i.e., ensembles comprised of successive model runs). In contrast, the Short-Range Ensemble Forecast System (SREF, horizontal grid spacing ~16 km) and Global Ensemble Forecast System (GEFS, effective horizontal grid spacing ~33 km) provide information on forecast uncertainty but fail to adequately resolve many key topographic features of the western U.S. As a result, meteorologists employ a variety of techniques to generate QPFs and PQPFs over the western U.S. using deterministic CPMs (Alexander et al. 2014; Rogers et al. 2017), ad-hoc ensembles

comprised of a collection of CPM forecasts (Alexander et al. 2011; Jirak et al. 2012, 2016), coarse-resolution ensembles, and downscaling approaches (Novak et al. 2014; Lewis et al. 2017).

The promise of CPEs over the western U.S. reflects their ability to both resolve fine-scale precipitation processes, including orographic effects, and estimate forecast uncertainty. The former reflects the ability of CPMs to produce precipitation forecasts with better-defined, more realistic precipitation structures than convection-parameterizing models (Mass et al. 2002; Roberts and Lean 2008; Weisman et al. 2008; Schwartz et al. 2009; Clark et al. 2015). For example, Roberts and Lean (2008) showed that forecasts of convective precipitation over the United Kingdom produced by the Met Office Unified Model (MetUM) at 1-km horizontal grid spacing without parameterized convection resulted in increased realism and skill compared to forecasts at 12-km grid spacing with parameterized convection. Similarly, Schwartz et al. (2009) found that QPFs of convection over the central United States produced by the Weather Research and Forecasting model (WRF) at 2-km horizontal grid spacing were more detailed than those produced by the WRF at 4-km grid spacing and superior to those generated by the operational 12-km NAM. In mountainous terrain, several studies have demonstrated that decreasing horizontal grid spacing to below 4-km improves simulations of orographic precipitation (Colle et al. 2005; Garvert et al. 2005; Schwartz 2014).

Ensembles produce estimates of forecast uncertainty by executing multiple model runs, each with varied initial conditions and/or model configurations. Because of their high resolution, CPEs can assess the inherent small-scale uncertainties at convective scales, which lead to rapid error growth (Lorenz 1969), and the sensitivity of orographic

precipitation to characteristics of the incident flow (Colle 2004; Roe 2005; Rotunno and Houze 2006). Using idealized simulations, Colle (2004) noted that the distribution and intensity of orographic precipitation is highly dependent on the speed of the incident flow, vertical wind shear, static stability, freezing level, and dimensions of the mountain barrier. Observational studies confirm these sensitivities and highlight the significance of low-level flow patterns (blocked or unblocked) on the distribution of orographic precipitation (Neiman et al. 2002; Stoelinga et al. 2003; Rotunno and Houze 2007; Smith et al. 2012).

Recent increases in computing capabilities in the U.S. have allowed for the assembling of operational, ad hoc CPEs such as the Storm Prediction Center Storm-Scale Ensemble of Opportunity [SSEO (Jirak et al. 2012, 2016)] and the High Resolution Rapid Refresh (HRRR) Time-Lagged Ensemble [HRRR-TLE (Alexander et al. 2011)] and the development of an experimental, but "true" (non-ad hoc), ensemble prediction system (EPS), the NCAR Ensemble [hereafter NCAR ENS (Schwartz et al. 2015)]. In Europe, several "true", operational CPEs have been developed including the Météo France Applications of Research to Operations at Mesoscale – Ensemble Prediction System [AROME-EPS (Bouttier et al. 2012; Vié et al. 2012)], the Deutscher Wetterdienst Consortium for Small-scale Modeling Ensemble Prediction System [COSMO-DE-EPS (Gebhardt et al. 2011)], and the Met Office Global and Regional Ensemble Prediction System [MOGREPS-UK (Tennant 2015)]. A key difference among these CPEs is the methods used to produce a set of forecasts. The SSEO uses a multimodel, multiphysics approach (Jirak et al. 2012), whereas the HRRR-TLE simply uses a series of time-lagged forecasts (Alexander et al. 2011). The NCAR ENS, AROME-EPS, and MOGREPS-UK utilize ensemble data assimilation to perturb the initial conditions (Bowler et al. 2008; Vié

et al. 2012; Schwartz et al. 2015), whereas nonstochastic physics perturbations are implemented in COSMO-DE-EPS (Gebhardt et al. 2011).

The majority of validation studies involving CPEs have focused on how different ensemble methods and model configurations affect their performance (e.g., Bouttier et al. 2012; Vié et al. 2012; Ben Bouallègue et al. 2013; Romine et al. 2014; Johnson and Wang 2016; Melhauser et al. 2017). Several have also investigated the ability of CPEs to forecast specific weather phenomena such as tornadoes (Gallo et al. 2016), convective initiation near the dryline (Trier et al. 2015), hurricanes (Zhang et al. 2010; Munsell et al. 2015), and stationary convective rain bands (Barrett et al. 2016). Although limited, studies comparing the warm-season QPF performance of CPEs to convection-parameterizing ensembles have largely produced promising results (Clark et al. 2009; Le Duc et al. 2013; Schellander-Gorgas et al. 2017). We are unaware of any cool-season QPF validation studies involving CPEs or any work intercomparing the performance of QPFs from CPEs, convection-parameterizing ensembles, and deterministic CPMs in any season.

This paper evaluates the performance of cool-season QPFs produced by the 3-km, 10-member, convection-permitting NCAR Ensemble (hereafter NCAR ENS) relative to several operational deterministic and probabilistic models at mountain locations over the western U.S. The high resolution of the NCAR ENS allows it to adequately resolve many key terrain features and their influence on precipitation, while also estimating forecast uncertainty. Because it is a single-physics, non-time-lagged CPE, unlike the SSEO and HRRR-TLE, each ensemble member is equally likely to represent the "truth", which allows for a more robust interpretation of probabilistic forecasts. Cool-season precipitation is validated because of the hazards it causes in the western U.S., such as flooding, avalanches,

and traffic and air accidents. In Chapter 2, we describe the models, datasets, and methods used in the paper, with key results and a model performance intercomparison presented in Chapter 3. The paper concludes with a summary, including a discussion of the significance of our findings for future model development and operational forecasting over the western U.S.

CHAPTER 2

DATA AND METHODS

2.1 NCAR Ensemble

Described in depth by Schwartz et al. (2015), the NCAR ENS produces forecasts for the conterminous U.S. and consists of an analysis component run at 15-km grid spacing and a 10-member forecast component run at 3-km grid spacing. Both the analysis and forecast components use version 3.6.1 of the Advanced Research WRF model (WRF-ARW) with 40 vertical levels and a parameterization suite that includes the Thompson microphysics scheme (Thompson et al. 2008), the Rapid Radiative Transfer Model for Global Climate Models (RRTMG) with ozone and aerosol climatologies for long- and short-wave radiation (Mlawer 1997; Tegen et al. 1997; Iacono et al. 2008), the Mellor-Yamada-Janić (MYJ) planetary boundary layer (PBL) scheme (Mellor and Yamada 1982; Janić 1994, 2002), and the Noah land surface model (Chen and Dudhia 2001). The analysis component also uses the Tiedtke cumulus parameterization (Tiedtke 1989). In the analysis component, an 80-member[1] continuously cycling ensemble adjustment Kalman filter (EAKF; Anderson 2001, 2003) produces analyses every 6 h (0000 UTC, 0600 UTC, 1200 UTC, and 1800 UTC). At 0000 UTC, the forecast component is initialized by interpolating

[1] The analysis component initially consisted of 50 members, but was upgraded to 80 members in May 2016.

10 members of the analysis component onto a 3-km grid nested within the 15-km domain. The smaller number of 3-km ensemble forecast members, compared to those in the EAKF system, reflects computational constraints. Nevertheless, 10 members are sufficient to produce skillful probabilistic forecasts (Clark et al. 2009, 2011; Schwartz et al. 2014). The forecast component then produces 48-h, 10-member, 3-km forecasts. For convenience, we refer to member 1 as the control member (hereafter NCAR ENS CTL). All NCAR ENS forecasts were obtained from NCAR's Research Data Archive (RDA).

2.2 Operational Models

We also examine the performance of several NCEP operational modeling systems including the HRRR, NAM-3km, Global Forecast System (GFS), and SREF. The SREF contains two dynamical cores, the WRF-ARW and the NCEP Non-hydrostatic Multiscale Model on the B grid (NMMB), each producing 13 ensemble members (Du et al. 2015). The control members of each core are referred to as the SREF ARW CTL and SREF NMMB CTL.

The most recent operational version of each model as of the end of the 2016/17 cool-season (31 March 2017) is used for the entirety of the validation period. In the case of the NAM-3km, which underwent a significant upgrade during the 2016/17 cool-season (Rogers et al. 2017), parallel, pre-operational runs are used prior to their operational implementation in mid-March, after which operational runs are used. HRRR and SREF forecasts were acquired from NCEP's NOAA Operational Model Archive and Distribution System (NOMADS). GFS forecasts and pre-operational forecasts from the NAM-3km were provided by NCEP Environmental Modeling Center. All modeling systems are

validated using output grids at their respective horizontal grid spacing. Table 2.1 provides a summary of basic information for each NCEP modeling system.

2.3 Precipitation Observations and Analyses

Gauge-based precipitation observations from the Snow Telemetry (SNOTEL) network are used to assess the performance of QPFs and PQPFs at mountain locations. SNOTEL sites are designed to collect snowpack, precipitation, and related climatic data. There are currently over 800 sites operated and maintained by the Natural Resources Conservation Service (NRCS). SNOTEL sites are typically located in sheltered locations that receive substantial snowfall. Precipitation is measured in large-storage gauges that measure hourly accumulated precipitation with a precision of 0.1 in. (~2.54 mm) using a manometer and pressure transducer (Serreze et al. 1999). Each gauge has a 30.5 cm orifice and an Alter wind shield to reduce undercatchment. Because of their sheltered locations, wind speeds at SNOTEL sites are generally less than 2 m s^{-1} (Ikeda et al. 2010). Nevertheless, undercatchment of ~10-15% has been shown for similar gauges under such conditions (Yang et al. 1988; Fassnacht 2004; Rasmussen et al. 2012) and likely artificially increases model biases in our results. Such undercatch is likely more significant at sites that are windier and receive lower density snow. Although the SNOTEL sites report hourly precipitation, we use only 24-h (1200-1200 UTC) accumulated precipitation totals to minimize the effect of artificial changes in the amount of reported precipitation as the ambient temperature fluctuates diurnally, causing the fluid in the precipitation gauges to expand and contract. Other issues that may affect SNOTEL precipitation data include transmission errors, instrument malfunction, and snow adhesion to the gauge walls. Owing

to these issues, we quality control the SNOTEL data following Lewis et al. (2017), resulting in data from 670 stations available for validation. Sites that had missing or erroneous data on 20% or more of the cool-season days were removed.

We also use daily (1200-1200 UTC) precipitation analyses produced by the PRISM Climate Group at Oregon State University (Daly et al. 1994, 2008; Luzio et al. 2008) to further illustrate spatial characteristics of model biases in selected mountainous regions. These daily analyses are available at 4-km grid spacing and are produced using observational point data, a digital elevation model, and spatial datasets (Daly et al. 1994).

2.4 Verification

Although forecasts by the NCAR ENS are available beginning in April 2015, we focus on the 2016/17 cool-season due to the availability of forecasts from the most recent versions of the NCEP operational models. Here, the 2016/17 cool-season is defined as 1 October 2016 through 31 March 2017. Each day, we validate 24-h QPFs ending at 1200 UTC on the day of interest. For example, January 25 refers to the 24-h period ending at 1200 UTC on January 25. We omitted days without precipitation forecasts from any modeling system from the study. Out of the 182 days in the 2016/17 cool-season, 28 days are omitted.

For all modeling systems except the HRRR and SREF, we perform validation using the 12-36-h QPFs initialized at 0000 UTC. Because the HRRR only provides forecasts to 18 h, we merge the 3–15-h QPFs from the forecasts initialized at 0900 and 2100 UTC to obtain an equivalent 24-h QPF. The SREF does not run at 0000 UTC, so we use the 9–33-h QPFs from forecasts initialized at 0300 UTC. Following Lewis et al. (2017), all model

QPFs are bilinearly interpolated to each SNOTEL site or PRISM grid point for calculations. Nearest neighbor interpolation was tested and produced nearly identical results.

A thorough evaluation of QPF requires an understanding of model biases and the analysis of several statistical verification measures (Schaefer 1990; Brill 2009). Following Mason (2003), we use statistical measures based on a standard 2x2 contingency table (Table 2.2) to evaluate deterministic forecasts including

$$Hit\ rate = \frac{a}{a+c} = \frac{hits}{observed\ events},$$

$$False\ alarm\ ratio = \frac{b}{a+b} = \frac{false\ alarms}{forecasted\ \ events},$$

and

$$Equitable\ threat\ score\ (ETS) = \frac{a - a_{ref}}{a - a_{ref} + b + c},$$

where

$$a_{ref} = \frac{(a+c) * (a+b)}{n}.$$

Hit rate measures the fraction of observed events correctly forecasted, false alarm ratio expresses the fraction of forecasted events that were false alarms, and ETS measures the fraction of observed and/or forecasted events that were correctly forecasted, adjusted for the frequency of hits expected by chance (climatology). While modern, convective-scale verification measures including 'neighborhood' approaches have been developed (e.g., Ebert 2008), we use the traditional, point-based ETS because cool-season precipitation in mountainous regions is strongly tied to terrain. Issues would arise using neighborhood approaches because of the dramatic changes in precipitation climatology over small spatial

scales that exist in mountainous regions.

We determine the quality of probabilistic forecasts from ensembles by computing their reliability and resolution, which are defined by:

$$Reliability = \frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \bar{o}_k)^2,$$

$$Resolution = \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2,$$

where N is the total number of forecasts, K is the total number of unique forecasts, $\bar{o}$ is the observed climatological frequency for the event to occur, n_k is the number of forecasts with the same probability, and $\bar{o}_k$ is the observed frequency of the event, given forecasts of probability f_k. Reliability assesses the statistical consistency between predicted probabilities and observed relative frequencies, whereas resolution measures the ability of an ensemble to distinguish when the event of interest occurs with lower or higher frequency than climatology. We also calculate the Brier Score (BS) for each ensemble, which measures the mean squared probability error and is given by:

$$BS = Reliability - Resolution + Uncertainty,$$

where

$$Uncertainty = \bar{o}(1 - \bar{o}).$$

Additionally, we measure the skill of the ensemble by computing the Brier Skill Score [BSS (Brier 1950; Murphy 1973; Wilks 2011)], which is defined as:

$$BSS = 1 - \frac{BS}{BS_{Cl}},$$

where BS_{Cl} is the BS of climatology. Good ensemble performance is indicated by lower values of BS and reliability and higher values of BSS and resolution. We use attribute

diagrams to visually assess these statistical measures and evaluate other ensemble characteristics (Toth et al. 2003). Consistency resampling (Brocker and Smith 2007) and bootstrap resampling (Efron and Tibshirani 1993; Hamill 1999) are employed to produce 5% and 95% consistency bars and confidence intervals, respectively, for the attribute diagrams, which involves resampling 1000 times and choosing N samples with replacement, where N is the total number of forecasts.

All of these measures require that the event of interest be dichotomous (yes/no). Therefore, we apply a threshold to each event, which we define as the total accumulated observed or forecast precipitation in a 24-h (1200 UTC to 1200 UTC) period (including 24-h periods with no precipitation). In addition to using absolute event thresholds (e.g., 15 mm, 20 mm, 25 mm, etc.), we use event percentile thresholds (e.g., 75th percentile, 80th percentile, 85th percentile, etc.). Following Roberts and Lean (2008) and Dey et al. (2014) we compute the distribution of events observed at SNOTEL sites and forecasted by each deterministic model and ensemble member to determine percentile thresholds for the observed and forecast events. Because we compare percentile thresholds from observed and forecast events, the absolute thresholds corresponding to a given percentile threshold for the observations and forecasts can differ. For example, the 95th percentile, which represents the top 5% of events, may be 35 mm for a certain model and 25 mm for SNOTEL observations. This method implicitly removes bias, allowing for an assessment of the spatial placement of precipitation within the context of each model's climatology, and reduces sampling issues resulting from differing observed and forecast precipitation climatologies across the western U.S.

Table 2.1. Characteristics of modeling systems and forecasts used in this study.

Forecast system	Acronym	Modeling center	Approximate horizontal grid spacing	Convective parameterization	No. of ensemble members	QPF used
NCAR Ensemble	NCAR ENS	NCAR	3-km	N/A	10	12-36 h from 0000 UTC
HRRRv2	HRRR	NCEP	3-km	N/A	N/A	03-15 h from 0900 and 2100 UTC
NAMv4 3-km CONUS nest	NAM-3km	NCEP	3-km	N/A	N/A	12-36 h from 0000 UTC
GFSv13.0.2	GFS	NCEP	13-km	simplified Arakawa-Schubert	N/A	12-36 h from 0000 UTC
SREFv7.0	SREF	NCEP	16-km	Multiple	26	09-33 h from 0300 UTC

Table 2.2. Contingency table used for validation.

		Observed		
		Yes	No	Total
Forecast	Yes	Hit (*a*)	False alarm (*b*)	*a* + *b*
	No	Miss (*c*)	Correct rejection (*d*)	*c* + *d*
	Total	*a* + *c*	*b* + *d*	*n*

CHAPTER 3

RESULTS

3.1 Observed and Forecast Cool-Season Precipitation Characteristics

3.1.1 Synopsis of 2016/17 Cool-Season Precipitation

Significant spatial variations in precipitation existed across the western U.S. during the 2016/17 cool-season. The cool-season was generally wetter than average across all of the western U.S., except for portions of Colorado, southern Utah, Arizona, and New Mexico, where precipitation was average to slightly below average (not shown). At upper elevations, mean daily precipitation ranged from > 16 mm in the Cascades and coastal ranges of the Pacific Northwest to < 3 mm in parts of the Rocky Mountains of Colorado and New Mexico, as well as other climatologically dry ranges of the western U.S. interior (Figure 3.1a,b). Measureable precipitation (≥ 2.54 mm[2]) occurred on ~70-80% of days in the Cascades and coastal ranges of the Pacific Northwest, ~45-70% of days in the Sierra Nevada and northern interior ranges, and ~25-45% of days in the southern interior ranges (Figure 3.2a,b). The magnitudes of 85th and 95th percentile events were generally greatest in the Cascades, coastal ranges from northern California to Washington, and Sierra Nevada, and decreased toward the interior ranges (Figure 3.3a-d). SNOTEL sites with relatively

[2] The precision of the precipitation gauges at SNOTEL sites is 0.1 in. (2.54 mm). Hence, the minimum amount of precipitation they can record is 2.54 mm.

large 85th and 95th percentile events in the interior were found in the Idaho Panhandle, Southwest Idaho Mountains, and the Mogollon Rim of Arizona (Figures 3.3a,c), regions that receive relatively large fractions of their climatological cool-season precipitation from inland penetrating atmospheric rivers (Rutz et al. 2014, 2015).

3.1.2 Model Biases

The ratio of forecast to observed mean daily precipitation (i.e., the bias ratio) identifies SNOTEL site locations where a model over (bias ratio > 1) or under (bias ratio < 1) predicts the total observed cool-season precipitation. Given undercatch and observational uncertainty, we consider bias ratios of 0.85-1.2 to be reflective of a near-neutral bias. For ensembles, we focus on the control member of each dynamical core. Therefore, the NCAR ENS has one (NCAR ENS CTL) and the SREF two control members (SREF ARW CTL and SREF NMMB CTL). Other members in each core exhibit similar bias ratios as their respective control runs, as will be shown in section 3.3. At SNOTEL sites, the NCAR ENS CTL produces a mean bias ratios ~1 with relatively low standard deviations of bias ratios at all SNOTEL sites, indicating each model's ability to accurately produce the total cool-season precipitation at mountain locations (Figures 3.4a). Aside from a dry bias at SNOTEL sites in Idaho and northwest Montana, the HRRR exhibits bias ratios similar to the NCAR ENS CTL (Figure 3.4b). The NAM-3km exhibits a large mean bias ratio of 1.319, indicative of a substantial wet bias (Figure 3.4c). Although the GFS and SREF ARW CTL also produce mean bias ratios of ~1, relatively high standard deviations (0.397 and 0.434, respectively) reflect sizeable dry or wet biases at individual SNOTEL sites (Figure 3.4d,e). In contrast, the SREF NMMB CTL has a significant dry

bias, especially in southern Utah and Colorado (Figure 3.4f).

Following Lewis et al. (2017), we divide the SNOTEL sites into two regions, Pacific ranges and interior ranges, that feature highly differentiated climatologies and terrain characteristics (Figure 3.5). Intermediate stations are not presented for brevity. Time series of accumulated precipitation averaged over all SNOTEL sites in each region provide information regarding regional model biases (Figure 3.6). The NCAR ENS CTL generated ~112% of the total observed precipitation over the Pacific ranges and about as much precipitation as observed by SNOTEL sites over the interior ranges. The HRRR produced only ~86% of the total observed precipitation in the interior ranges, reflective of a dry bias, but agreed more closely with observations in the Pacific ranges. Total precipitation produced by the GFS was close to observed in both regions. The NAM-3km produced excessive precipitation in both regions, especially over the interior ranges where it produced ~130% of the total observed precipitation. The SREF ARW CTL's total predicted precipitation was slightly greater than observed in both regions, while the SREF NMMB CTL produced the least total precipitation in both regions, including only ~78% of total observed precipitation over the Pacific ranges. Overall, these results are consistent with Figure 3.4.

Bias ratios computed relative to PRISM analyses illustrate some of the spatial characteristics of forecast precipitation over the western U.S. For brevity, we focus on bias ratios over the complex terrain surrounding Salt Lake City, Utah (SLC) and Lake Tahoe, California. In the region surrounding SLC, bias ratios produced by the NCAR ENS CTL, HRRR, and NAM-3km generally increase from west (windward side) to east (leeward side) across the Stansbury Mountains, Oquirrh Mountains, and Wasatch Range (Figures 3.7a-c).

The NCAR ENS and HRRR, for example, produce bias ratios < 1 on the western slopes, ~ 1 near the crests, and > 1 on the eastern slopes of these mountain ranges (Figures 3.7a,b). Although the NAM-3km has a wet bias over the eastern and western slopes of all three ranges, its local bias ratio maxima are on the eastern slopes, consistent with a bias ratio increase from west to east (Figure 3.7c). Despite poorly resolving the three ranges, the GFS also exhibits a general tendency for bias ratio to increase from the windward to leeward slopes (Figure 3.7d). The SREF ARW CTL and SREF NMMB CTL overpredict valley precipitation and underpredict mountain precipitation (Figures 3.7e,f).

In the region surrounding Lake Tahoe, bias ratios produced by the NCAR ENS CTL, HRRR, and NAM-3km similarly increase from west to east across the Sierra Crest, Carson Range, and Pine Nut Mountains, with all three models exhibiting pronounced wet biases on their eastern (leeward) slopes (Figures 3.8a-c). Bias ratios for the GFS, SREF ARW CTL, and SREF NMMB CTL exhibit minimal topographic dependence over the Sierra Crest and are generally < 1 over the Carson Range and Pine Nut Mountains (Figures 3.8d-f).

Overall, bias ratios computed relative to PRISM analyses broadly represent spatial bias ratio characteristics across the west. Although the mean bias ratio varies between regions and models, NCAR ENS, HRRR, and NAM-3km bias ratios typically increase as one moves climatologically downstream across mountain barriers. This could reflect a systematic bias in these modeling systems or biases in the PRISM analysis methods. If this reflects a model bias, it may be the result of poorly resolved orographic processes due to terrain smoothing or deficiencies in microphysical parameterizations that cause too much precipitation to be advected over mountain crests. Aside from a dry bias over very

narrow mountain ranges (i.e., Carson Range), spatial bias ratio characteristics in the lower-resolution GFS, SREF ARW CTL, and SREF NMMB CTL are less generalizable, likely because very narrow mountain ranges are not resolved and wider mountain ranges are inadequately represented at horizontal grid spacings of ≥ 13 km.

Next, we bin events (2.54-mm intervals) to examine the ratio of forecast to SNOTEL-observed event frequencies (i.e., frequency bias) as a function of event size (Figure 3.9). We assume frequency biases > 1.2 reflect a clear overprediction of event frequency and < 0.85 a clear underprediction. Except for the NAM-3km, which overpredicts events > 36 mm, and SREF NMMB CTL, which underpredicts events < 30 mm, all models generally exhibit frequency biases between 0.85 and 1.2 for all event sizes in the Pacific Ranges (Figure 3.9a). Aside from the HRRR, frequency bias scores are generally worse over the interior ranges (Figure 3.9b). The NCAR ENS CTL overpredicts events > 28 mm and the NAM-3km overpredicts events > 18 mm. The NAM-3km overprediction grows nearly monotonically with event size, with a frequency bias > 2 for events > 39 mm. The GFS exhibits better frequency biases than the NCAR ENS CTL and NAM-3km, but the GFS underpredicts events > 42 mm. Except for an overprediction of events > 42 mm, the SREF ARW CTL generally displays no clear signs of overprediction or underprediction. The SREF NMMB CTL significantly underforecasts the frequency of events < 22 mm and overforecasts the frequency of events > 38 mm.

Overall, we find the least bias present in the NCAR ENS CTL and HRRR. Both models produce accurate cool-season precipitation totals at most SNOTEL sites. A slight wet bias in the NCAR ENS CTL and dry bias in the HRRR is revealed when looking at total precipitation averaged over both regions. Aside from the NCAR ENS CTL producing

too many large events in the interior ranges, both models generate an accurate number of events. Conversely, the NAM-3km exhibits a significant wet bias at most SNOTEL sites, while the GFS and SREF ARW CTL have minimal bias for cumulative SNOTEL site statistics, but a substantial wet or dry bias from site to site. A dry bias, due to too few small and moderate events, is found in the SREF NMMB CTL.

3.1.3 Distributions of Forecasted Events

We now focus on forecasts and their corresponding observations (i.e., event pairs) using bivariate histograms (Figure 3.10). More frequent event pairs falling near the 1-to-1 line with minimal skewness reflects low bias and good correspondence between forecast and observed events, while frequent event pairs above (below) the 1-to-1 line reflects underprediction (overprediction). Large scatter and a relatively large distance between conditional forecast and observed median values for the same event size reflects poor correspondence between forecast and observed events. In both regions, the NCAR ENS CTL has minimal skewness and moderate scatter (Figures 3.10a,c). It displays reasonable accuracy with slightly larger scatter in the interior ranges. Aside from slight skewness above the 1:1 line for events < 20 mm (underprediction) in the Pacific ranges, the HRRR displays high accuracy and minimal bias in both ranges (minimal skewness and scatter; Figures 3.10b,d). Consistent with its previously discussed wet bias, the NAM-3km is heavily skewed below the 1:1 line for all event sizes in both regions, indicating a tendency to overpredict (Figure 3.10e,g). Other than skewness above the 1:1 line for events > 10 mm in the Pacific ranges, the GFS displays accuracy similar to the NCAR ENS CTL (Figures 3.10f,h). The SREF ARW CTL displays substantial scatter in both regions and a

skewness below the 1:1 line in the Pacific ranges for events < 22 mm, indicating overprediction (Figures 3.10i,k). Very poor performance is shown by the SREF NMMB CTL with scatter so large that there appears to be no correlation between forecasts and observations (Figure 3.10j,l). Overall, we find that the skewness present in each model is generally consistent with previously discussed biases. Considering that minimal skewness and scatter indicate high accuracy, we determine the HRRR to be most accurate, followed by the NCAR ENS CTL, GFS, and NAM-3km, all with similar accuracy, then the SREF ARW CTL, and lastly the SREF NMMB CTL with minimal accuracy.

3.2 Determinitic Accuracy Measures

We now evaluate statistical measures based on a standard 2x2 contingency using absolute event thresholds to determine model performance characteristics as a function of event size. Aided by its wet bias, the NAM-3km scores the highest hit rates over both regions for all event thresholds (> 0.6 over Pacific ranges and ≥ 0.4 over interior ranges; Figure 3.11a,b). The NCAR ENS CTL, HRRR, GFS, and SREF ARW CTL produce similar hit rates for event thresholds < 23 mm, while the NCAR ENS CTL and HRRR score slightly higher than the GFS and SREF ARW CTL for event thresholds > 23 mm over the Pacific ranges. Over the interior ranges, the NCAR ENS CTL's hit rate improves relative to other models and is greater than or equal to the HRRR's for all event thresholds (Figure 3.11b). The hit rate for the GFS and SREF ARW CTL drop off considerably for event thresholds > 23 mm over the interior ranges. The SREF NMMB CTL performs poorly in both regions, recording hit rates < 0.5 for all event thresholds (Figure 3.11a,b). The HRRR produces the lowest false alarm ratios for all thresholds in both the Pacific and interior

ranges (Figure 3.11c,d). Again, we find a substantial improvement in the NCAR ENS CTL's scores over the interior ranges compared to the Pacific ranges (Figure 3.11c,d); its false alarm ratio is relatively poor (> 0.4 for event thresholds > 20 mm) and similar to that of the NAM-3km and SREF ARW CTL over the Pacific ranges, but improves relative to all other models and is similar to that of the GFS over the interior ranges (Figures 3.11c,d). Even with its significant dry bias, the SREF NMMB CTL records the worst false alarm ratios for all event thresholds over both regions (Figures 3.11c,d).

Over both the Pacific and interior ranges, the HRRR and NAM-3km generally produce the highest ETSs (Figure 3.11e,f). Because models with larger biases tend to have higher ETS (Mason 1989), the NAM-3km's ETS is likely aided by its wet bias. The GFS is more skillful (larger ETSs) than the NCAR ENS CTL over the Pacific ranges, but is less skillful (smaller ETSs) over the interior ranges. Consistent with other statistical measures, the SREF ARW CTL and especially the SREF NMMB CTL exhibit minimal skill over both ranges (Figures 3.11e,f). A general decline in ETSs by all models is evident over the interior ranges, especially for event thresholds > 25 mm. Overall, the highest resolution deterministic models perform best, as they are able to most fully resolve the terrain and thus orographic precipitation. The NCAR ENS CTL may have less skill relative to all other models over the Pacific ranges compared to the interior ranges because the western boundary of its 3-km forecast domain is very close to the Pacific coast (Schwartz et al. 2015). The western domains of other regional models are much further from the Pacific coast [HRRR (Alexander et al. 2014); NAM-3km (Carley et al. 2017); SREF (Du et al. 2015)].

We now focus on the same deterministic statistical measures using upper-quartile

and greater percentile event thresholds to evaluate bias-corrected model performance. Percentiles computed from SNOTEL observed and forecast events used to validate percentile event thresholds reveal model biases consistent with previous results (Figure 3.12). In general, bias correction improves the hit rate of models with a dry bias (i.e., the HRRR) and reduces the hit rate of models with a wet bias (i.e., the NAM-3km). Therefore, the HRRR exhibits the highest hit rates in both regions, followed by the NAM-3km and GFS in the Pacific ranges and the NAM-3km and NCAR ENS CTL in the interior ranges (Figures 3.13a,b). Contrary to the effect of bias correction on hit rates, false alarm ratios worsen (increase) for models with a dry bias and improve (decrease) for models with a wet bias when bias correction is applied (Figures 3.13c,d). The impact of removing bias on ETS is subtler, but we do find slight improvements in the scores of models with a dry bias and slight declines in the scores of models with a wet bias, such that the HRRR produces higher ETSs than the NAM-3km over both regions for almost all thresholds (Figures 3.13e,f). Overall, we find the bias-corrected results (Figure 3.13) to be generally consistent with the non-bias-corrected results (Figure 3.11) when accounting for the impact that bias has on these three statistical measures.

3.3 Probabilistic Verification

Similar to the method used for bias-corrected, deterministic validation, we now focus on the quality of PQPFs from the NCAR ENS and SREF using percentile event thresholds. Ideally, each member of an ensemble should be equally likely to be correct and, thus, all members should have identical climatologies. A tight packing of precipitation distributions for each member of the NCAR ENS reveals that each member indeed contains

a similar climatology, confirming the expectation of equal likelihood due to EAKF initializations. The climatologies of its members are characterized by a wet bias for 80th percentile and larger events in both regions (Figure 3.14). Conversely, an exceptional bifurcation is present in the distributions of SREF members. Clearly, the design of the SREF violates the principal of equal likelihood. Its use of two dynamical cores results in the distinct bifurcation of the distributions, while its use of different physics within each core generates greater spread among the distributions within the two clusters compared to the NCAR ENS. While the SREF ARW members contain a wet bias, the SREF NMMB members exhibit a sizeable dry bias, especially for 85th percentile events and smaller (Figure 3.14). Because of the dramatic differences in the climatologies of the two SREF cores we focus on the performance of the individual cores, in addition to the entire, 26-member SREF.

We use attributes diagrams to determine the probabilistic performance of the NCAR ENS, SREF, and the individual dynamical cores of the SREF (SREF ARW and the SREF NMMB) at forecasting 85th and 95th percentile events. Attributes diagrams provide information regarding the Brier score decomposition (reliability and resolution) and other characteristics of each ensemble. The shape of the reliability curves for the SREF and especially the NCAR ENS for 85th percentile events in both regions display overconfidence (Figure 3.15). For example, over the Pacific ranges, when the NCAR ENS forecasts a 90% probability that an 85th percentile event will occur, it only occurs ~67% of the time (Figure 3.15a). Similarly, when it forecasts a 10% probability that the event will occur, it occurs ~24% of the time. The SREF has better reliability in both regions and better resolution over the Pacific ranges, leading to higher BSSs (0.349 over the Pacific ranges and 0.318

over the interior ranges) than the NCAR ENS (0.296 over the Pacific ranges and 0.314 over the interior ranges; Figure 3.15). Overlap in confidence intervals for about half of the plotted points in each region decreases the significance of these results, especially over the Pacific ranges where differences in performance measures are minimal. Although BSS equally weighs the reliability and resolution (i.e., ability to distinguish when the event of interest occurs with lower or higher frequency than climatology) in determining its skill, resolution is considered the most important attribute of an ensemble (Toth et al. 2003). While reliability can be increased using *a posterior* calibration techniques, resolution cannot and can only be improved by a clearer segregation of scenarios where the event of interest occurs with higher or lower frequency than climatology (i.e., a better forecast in a probabilistic sense). While sharpness, which measures specificity of a probabilistic forecast, is not a measure of accuracy because it is only a function of the forecast, good sharpness is desirable in conjunction with strong reliability (Murphy 1993). The forecast frequency histograms reveal that the NCAR ENS forecasts high or low probabilities more often than the SREF, indicating greater sharpness (Figure 3.15). However, overconfidence and relatively poor reliability indicate that the NCAR ENS is likely too sharp (spread deficient).

We find similar performance characteristics in the NCAR ENS and SREF when focusing on 95th percentile event thresholds (Figure 3.16). Overconfidence is again evident in both ensembles, although to a lesser extent. While SREF continues to outperform the NCAR ENS over the Pacific ranges, with better reliability and resolution (Figure 16a), the NCAR ENS produces a larger BSS over the interior, aided by good resolution (Figure 3.16b). The NCAR ENS forecasts probabilities of 1 more than twice as much as the SREF

over the interior ranges, indicating more sharpness (Figure 3.16b). Although the SREF has a much coarser horizontal grid spacing (16-km) than the NCAR ENS (3-km), its PQPFs are often more skillful. While the NCAR ENS is too sharp with relatively poor reliability, the SREF contains more spread, largely due to its two climatologically contrasting dynamical cores, leading to less overconfidence. In other words, the SREF contains more spread because it violates the principal of equal likelihood.

Evaluating the performance and characteristics of the two, 13-member SREF cores (SREF ARW and SREF NMMB) provides insights into reasons for the characteristics of the full, 26-member SREF. Under all scenarios (85^{th} and 95^{th} percentile event thresholds in both regions), the SREF NMMB exhibits better reliability and resolution and hence larger BSSs (Figures 3.17 and 3.18). The SREF ARW suffers from significant overconfidence under all scenarios. Frequency histograms reveal a lack of sharpness (large spread) in the SREF NMMB, especially over the interior ranges for 85^{th} and 95^{th} percentile event thresholds (Figures 3.17b and 3.18b). Given that one would not expect an individual member of an ensemble that violates the principal of equal likelihood to perform well deterministically, this corresponds well with the dismal performance of the SREF NMMB CTL. The contrasting performance characteristics of the SREF ARW (poor reliability and resolution, reasonable sharpness) and SREF NMMB (good reliability and resolution, minimal sharpness), along with their differences in climatology, create an often more skillful probabilistic forecast than they would individually (Figures 3.15, 3.16, 3.17, and 3.18; see BSSs). Although the 26-member SREF is generally more skillful than the NCAR ENS, the NCAR ENS often outperforms the individual SREF dynamical cores; the NCAR

ENS is more skillful than the SREF ARW over the entire western US and the SREF NMMB over the interior ranges (Figures 3.15, 3.16, 3.17, and 3.18; see BSSs).

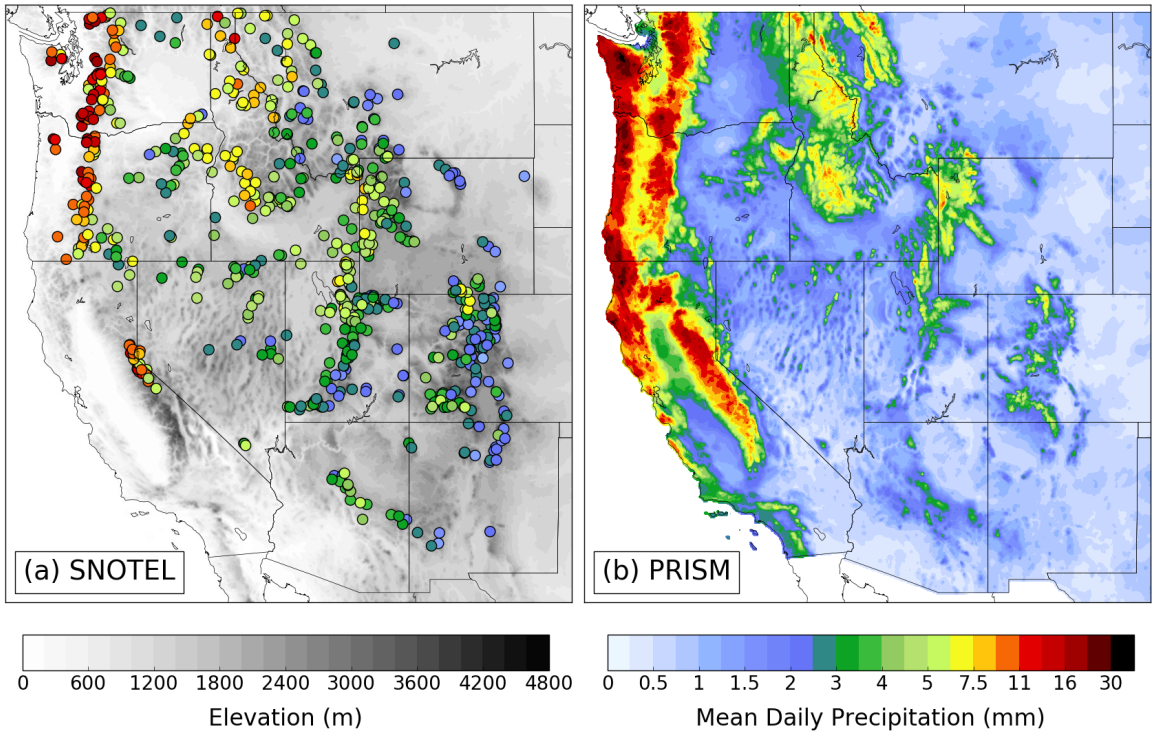


Figure 3.1. Mean daily precipitation. (a) Observed at SNOTEL sites (mm, color scale at bottom right) with 30 arc-second topography (m MSL, grey-shade scale at bottom left) and (b) analyzed by PRISM [mm, color shaded as in (a)].

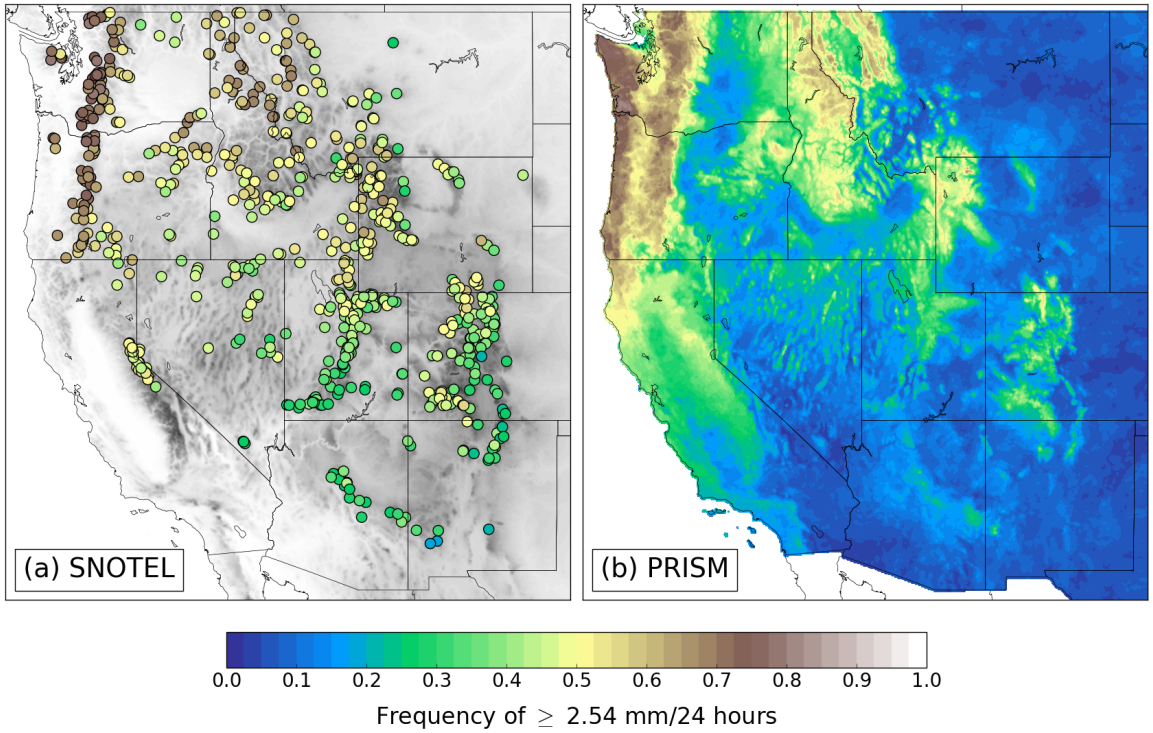


Figure 3.2. Frequency of precipitation events (≥ 2.54 mm). (a) Observed by SNOTEL sites (color scale at bottom) with 30-arc second topography (as in Figure 3.1a) and (b) analyzed by PRISM [as in (a)].

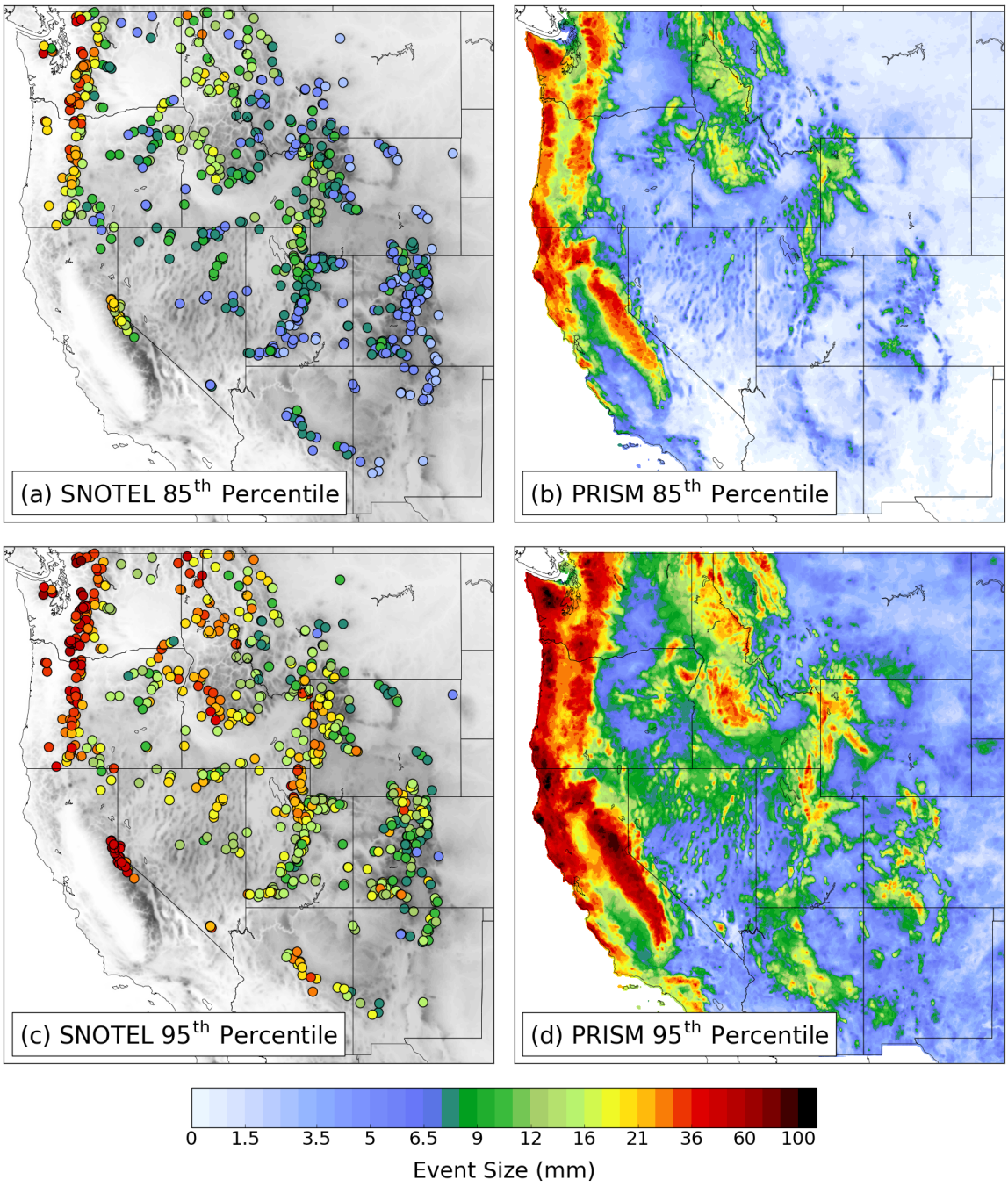


Figure 3.3. Magnitude of precipitation events. (a) 85th percentile events at SNOTEL sites (mm, color scale at bottom) with 30-arc second topography (as in Figure 3.1a). (b) 85th percentile events from PRISM analyses [as in (a)]. (c), (d) As in (a), (b), but for 95th percentile events.

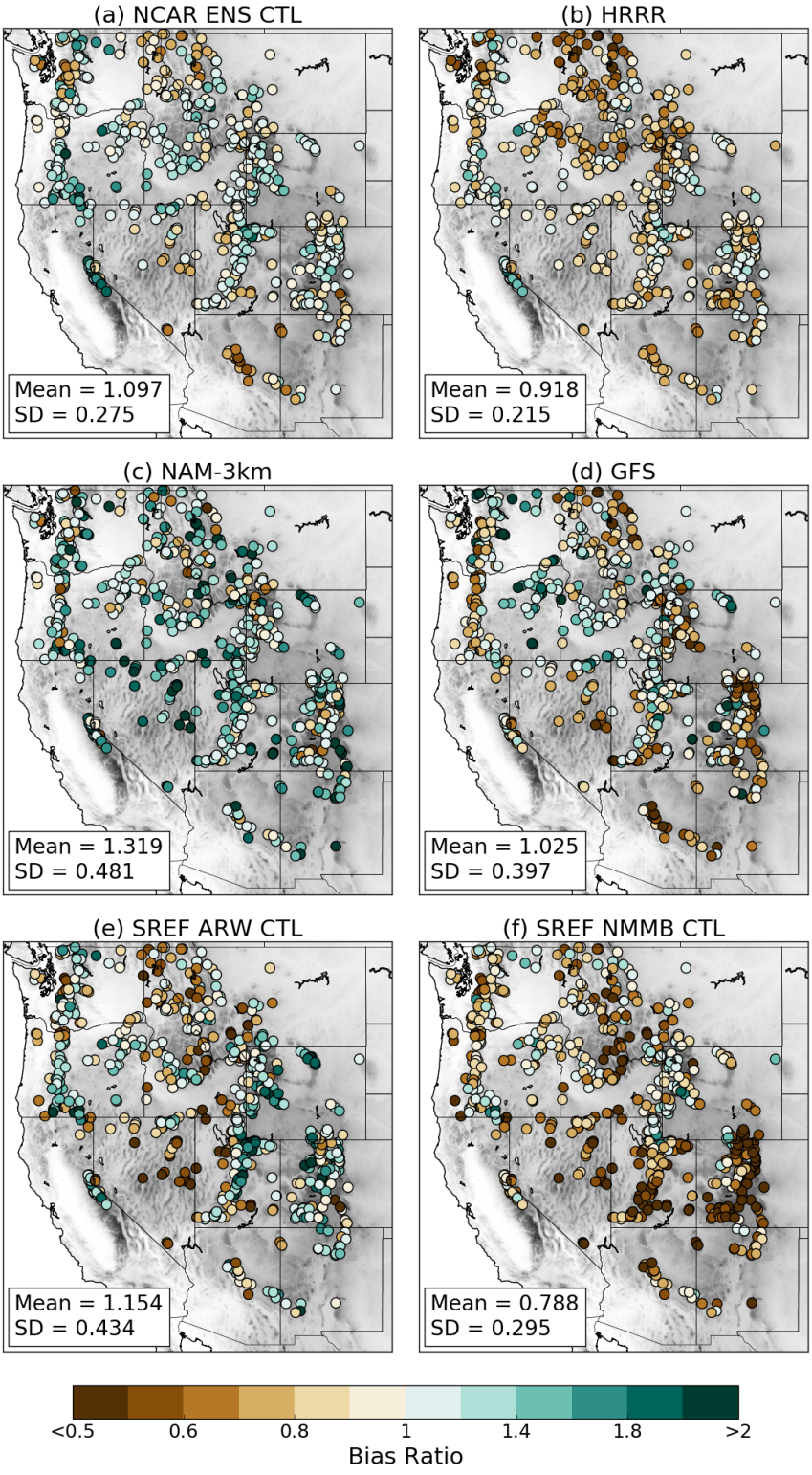


Figure 3.4. Bias ratios at SNOTEL sites (color scale at bottom) and 30-arc second topography (as in Figure 3.1a) with mean bias ratio and standard deviation (SD) annotated. (a) NCAR ENS CTL. (b) HRRR. (c) NAM-3km. (d) GFS. (e) SREF ARW CTL. (f) SREF NMMB CTL.

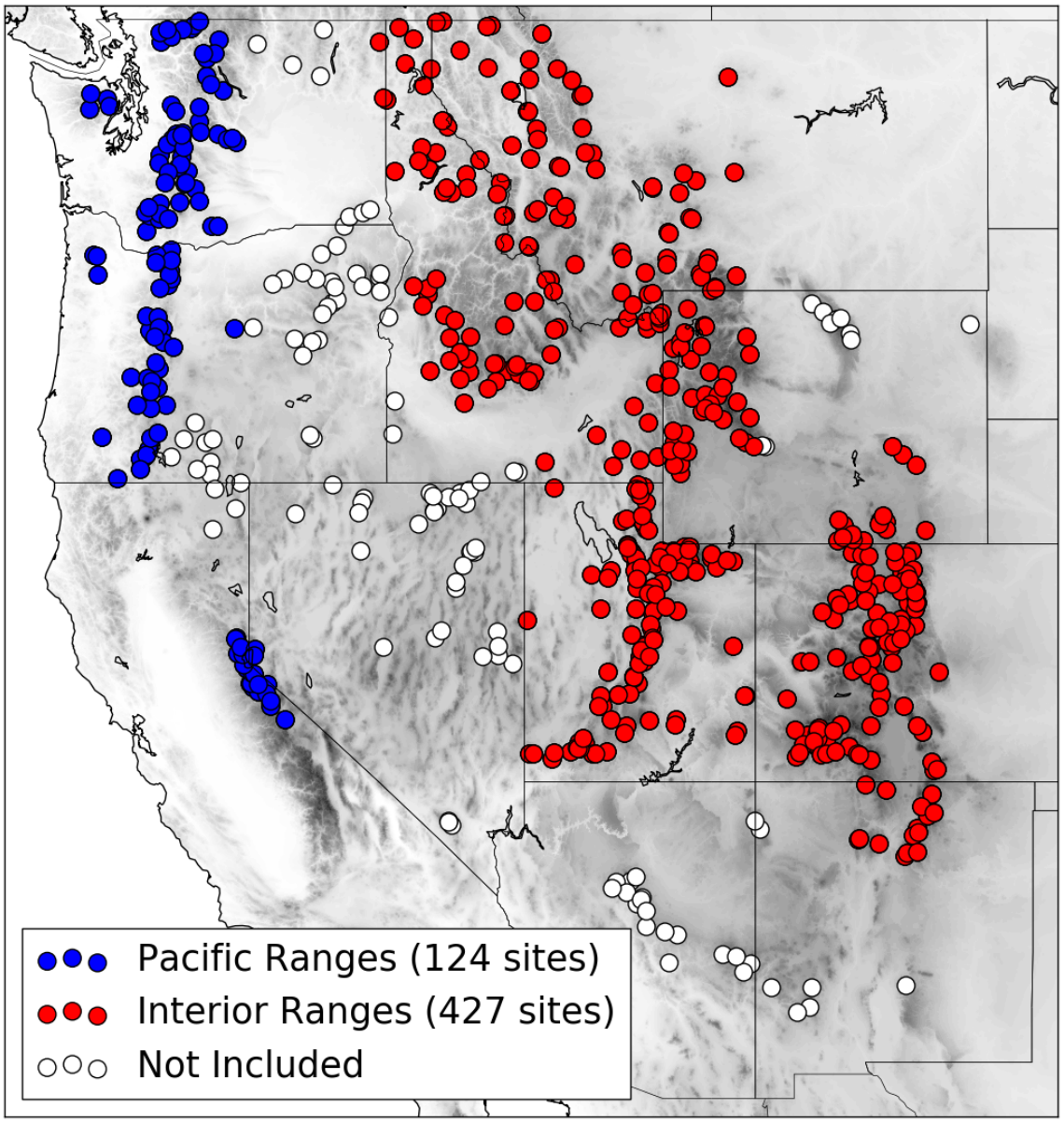


Figure 3.5. Regional classification of SNOTEL sites and 30-arc second topography (as in Figure 3.1a)

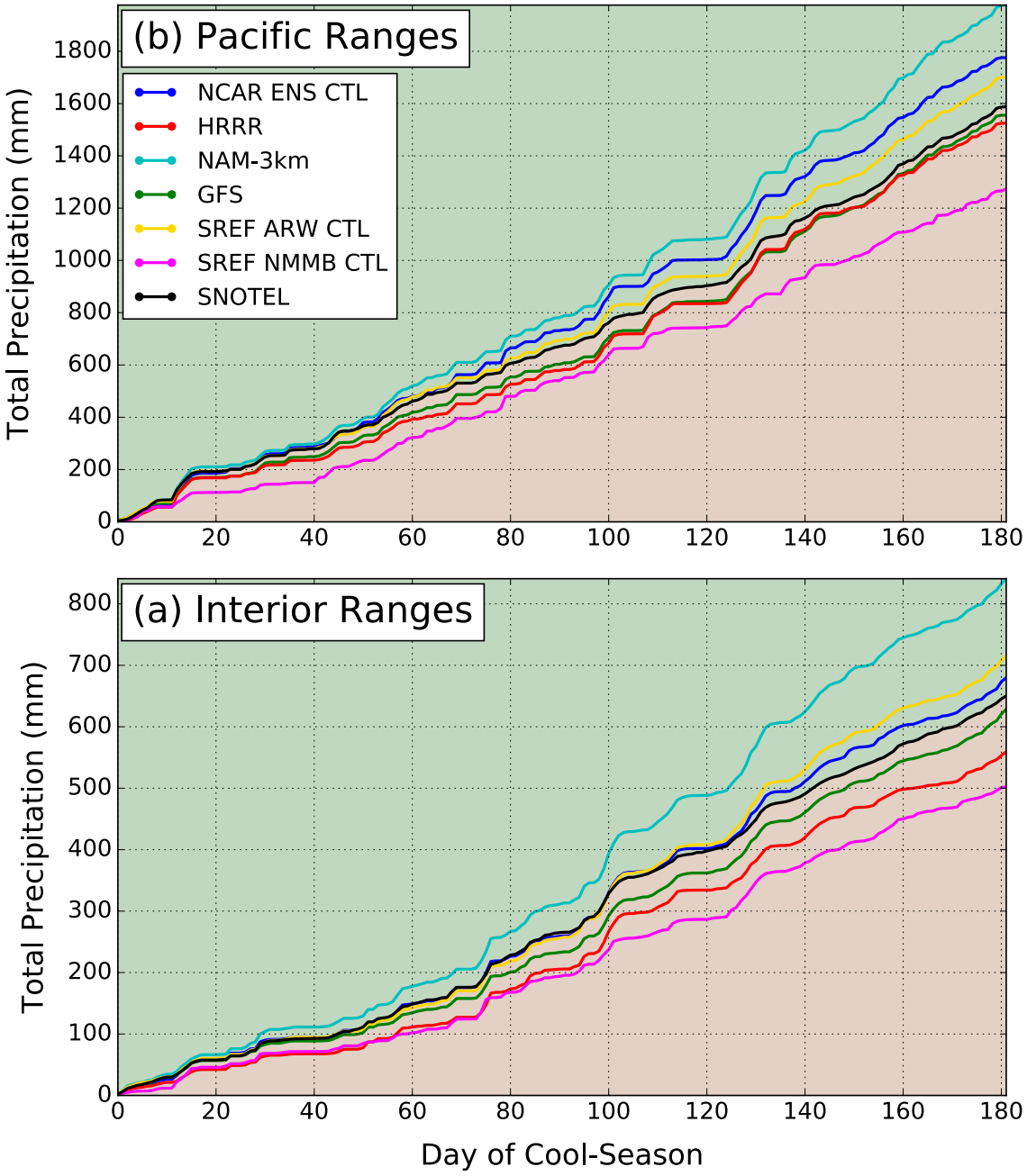


Figure 3.6. Mean observed and forecast accumulated cool-season precipitation at SNOTEL sites in the (a) Pacific ranges and (b) interior ranges. Light green (light brown) shading indicates above (below) the SNOTEL mean.

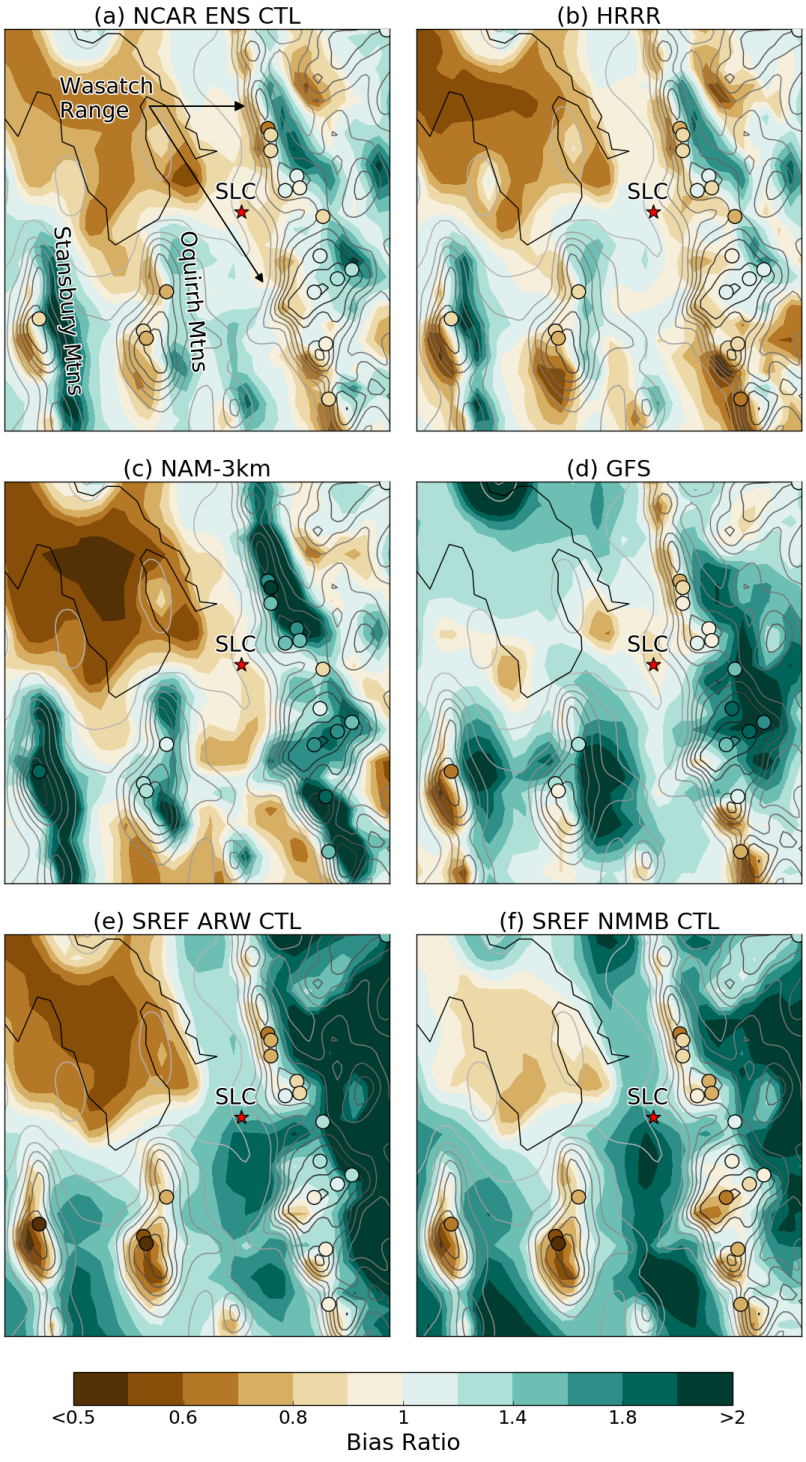


Figure 3.7. Bias ratios relative to PRISM analyses (following scale at bottom) and SNOTEL observations (filled circles following scale at bottom) in the region surrounding SLC for the (a) NCAR ENS CTL, (b) HRRR, (c) NAM-3km, (d) GFS, (e) SREF ARW CTL, and (f) SREF NMMB CTL. 1 arc-minute topography smoothed using a rectangular smoother and contoured every 200 m from 1300 m MSL (light grey) to 3300 m MSL (black). Mountain ranges annotated in (a).

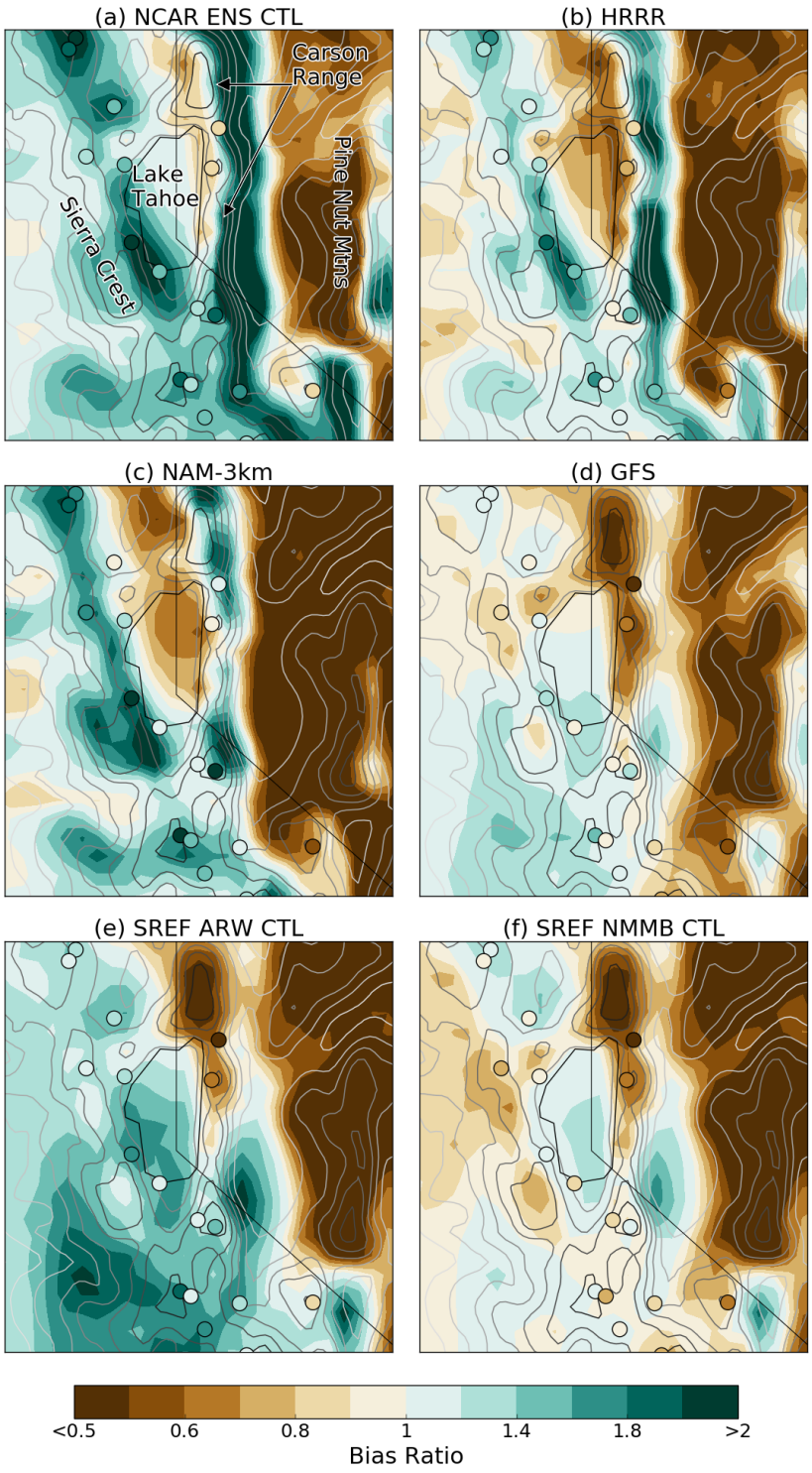


Figure 3.8. Same as Figure 3.7 except for the Lake Tahoe Region and topography contoured every 200 m from 1000 m MSL (light grey) to 2800 m MSL (black). Lake Tahoe and mountain ranges are annotated for reference in (a).

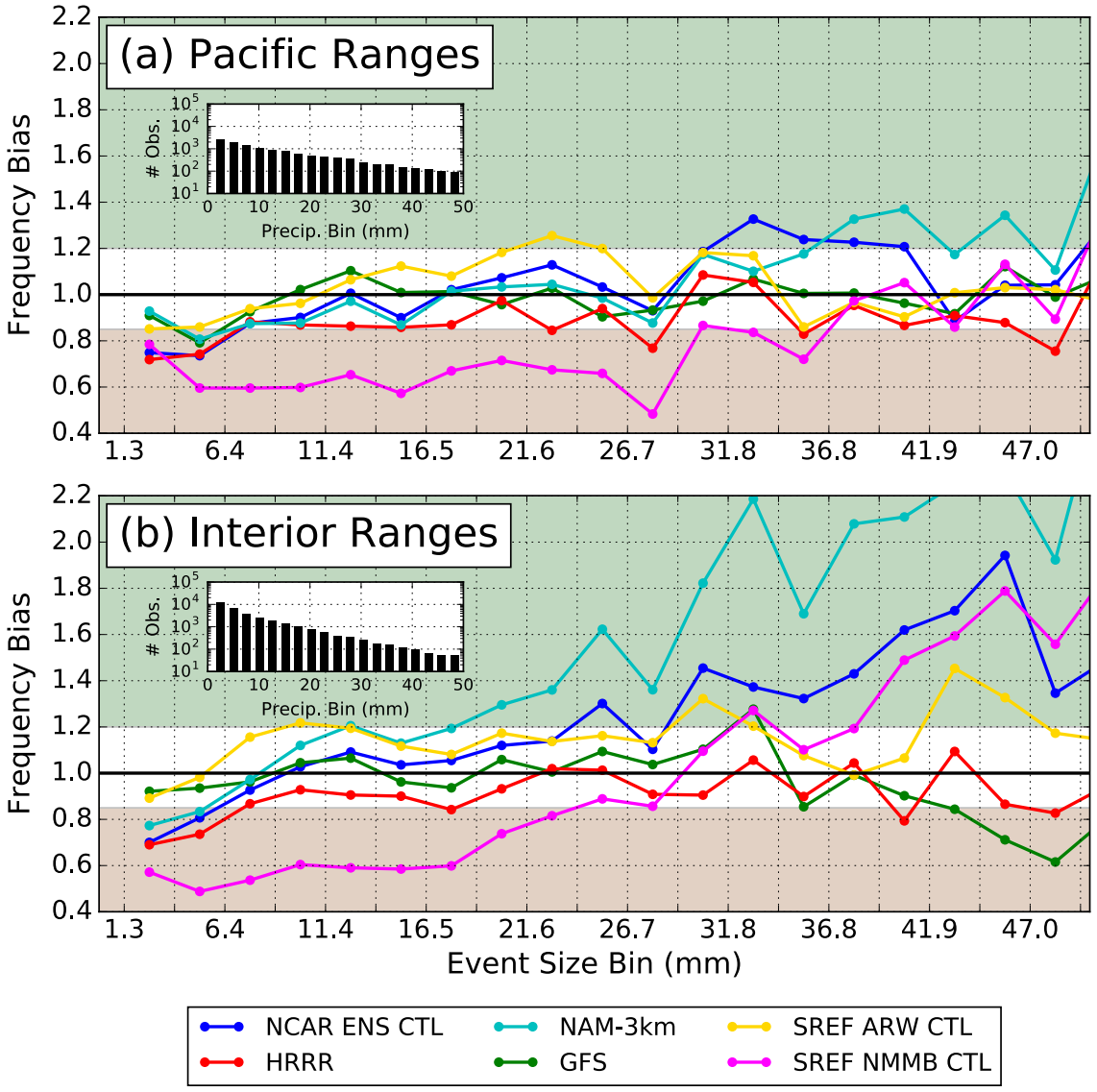


Figure 3.9. Frequency bias as a function of event size at SNOTEL sites in the (a) Pacific ranges and (b) interior ranges. Green (brown) shading indicates bias ratios ≥ 1.2 (≤ 0.85). Samples size in each bin shown in inset histograms.

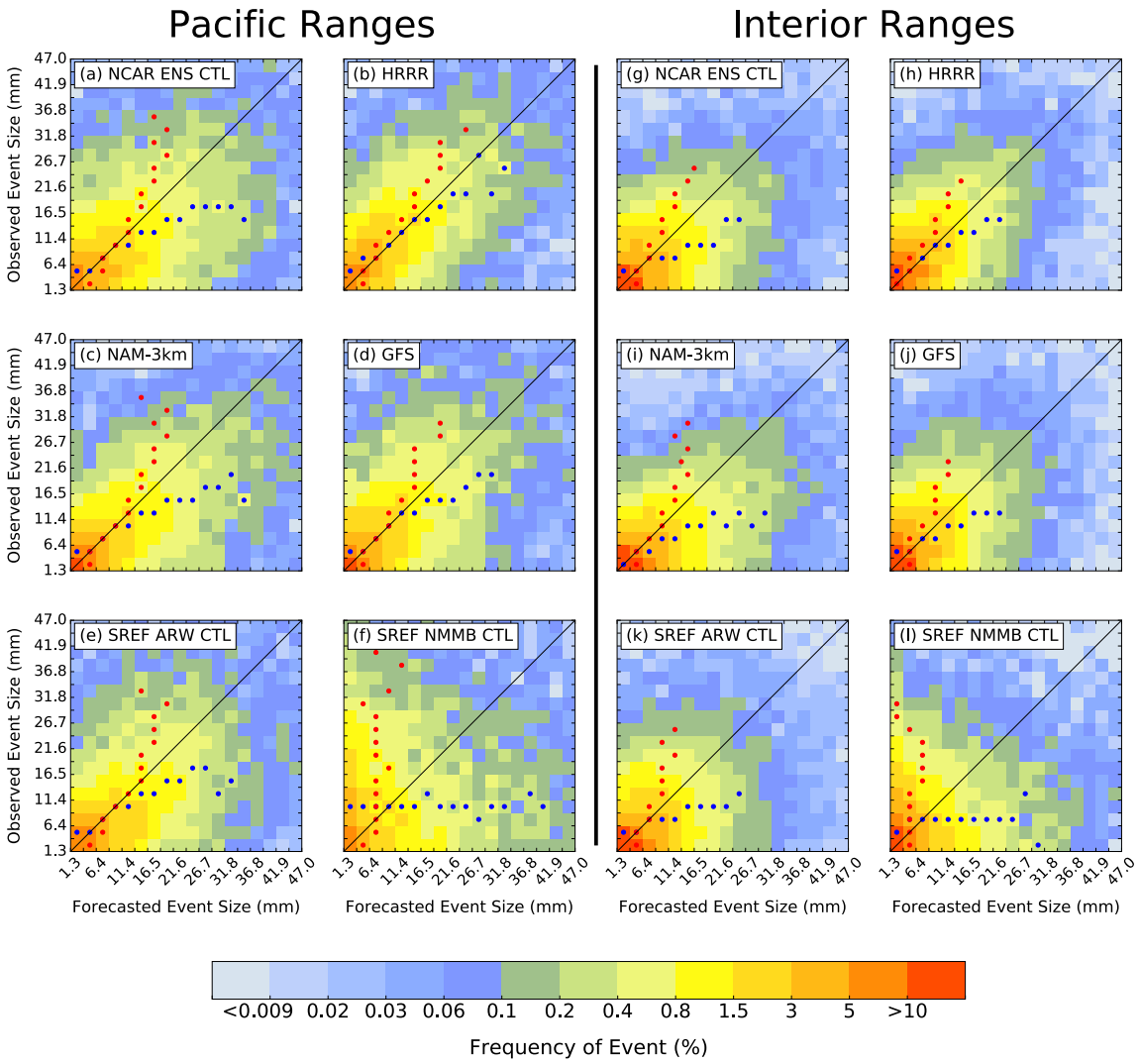


Figure 3.10. Bivariate histograms of forecast and observed precipitation at SNOTEL sites in the Pacific ranges for the (a) NCAR ENS CTL, (b) HRRR, (c) NAM-3km, (d) GFS, (e) SREF ARW CTL, and (f) SREF NMMB CTL. (g), (h), (i), (j), (k), (l) As in (a), (b), (c), (d), (e), (f), but over the interior ranges. Red (blue) dots represent the median observed (forecast) event size in each bin. Dots are not shown for bins with fewer than 50 events.

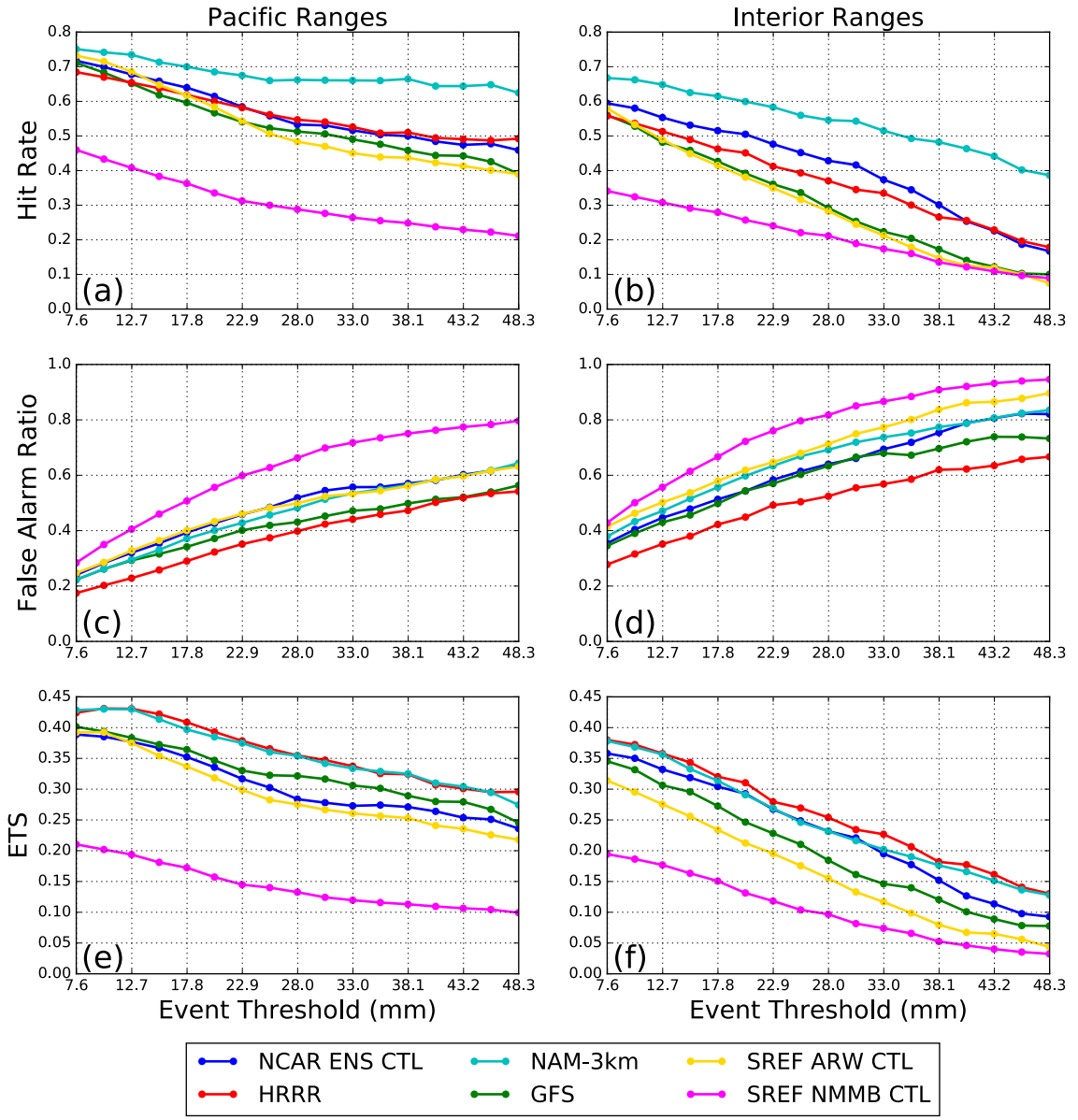


Figure 3.11. Verification metrics based on Table 2.2 as a function of absolute event thresholds (mm) at SNOTEL sites. (a) Hit rate in the Pacific ranges. (b) Hit rate in the interior ranges. (c), (d) Same as (a), (b) except False Alarm Ratio. (e), (f) Same as (a), (b) except Equitable Threat Score (ETS).

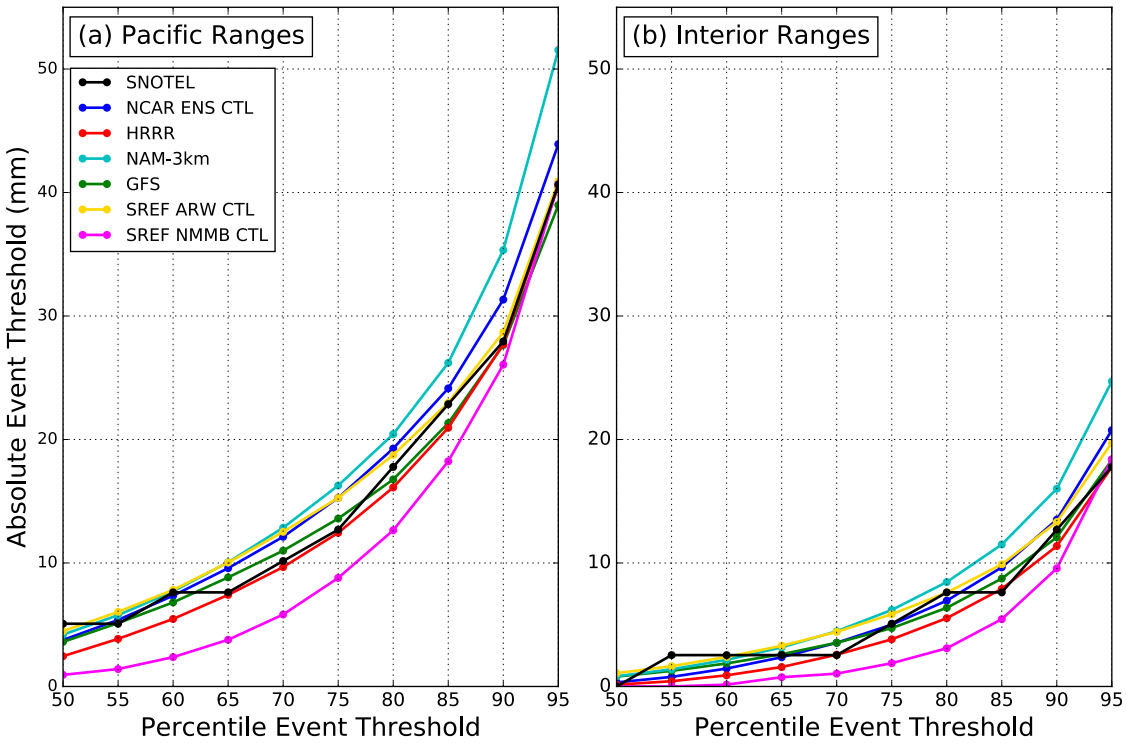


Figure 3.12. Forecast and observed absolute event thresholds (mm) corresponding to percentile thresholds for all forecasted and observed events at SNOTEL sites in the (a) Pacific ranges and (b) interior ranges.

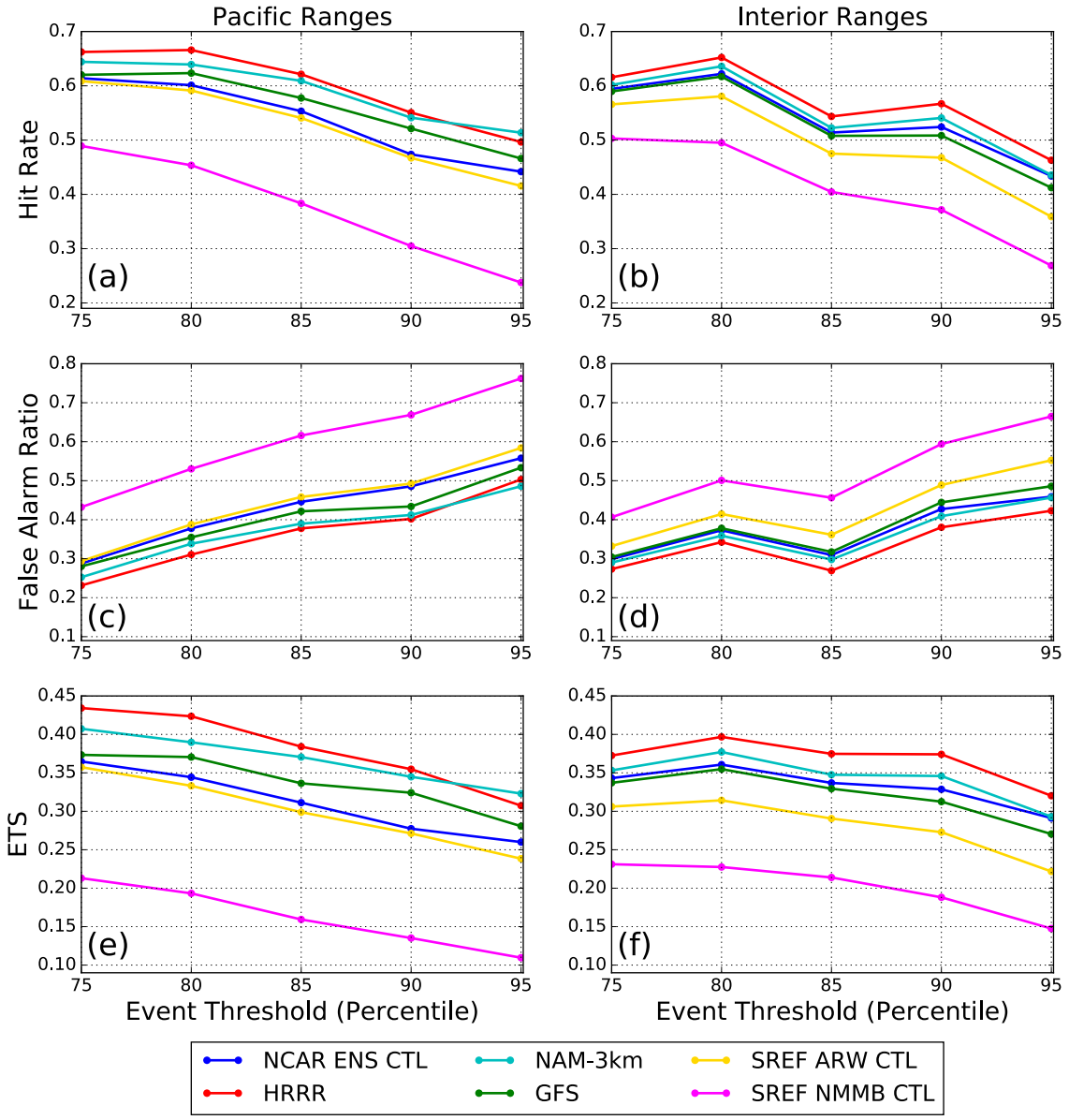


Figure 3.13. Same as Figure 3.11 except based on percentile event thresholds.

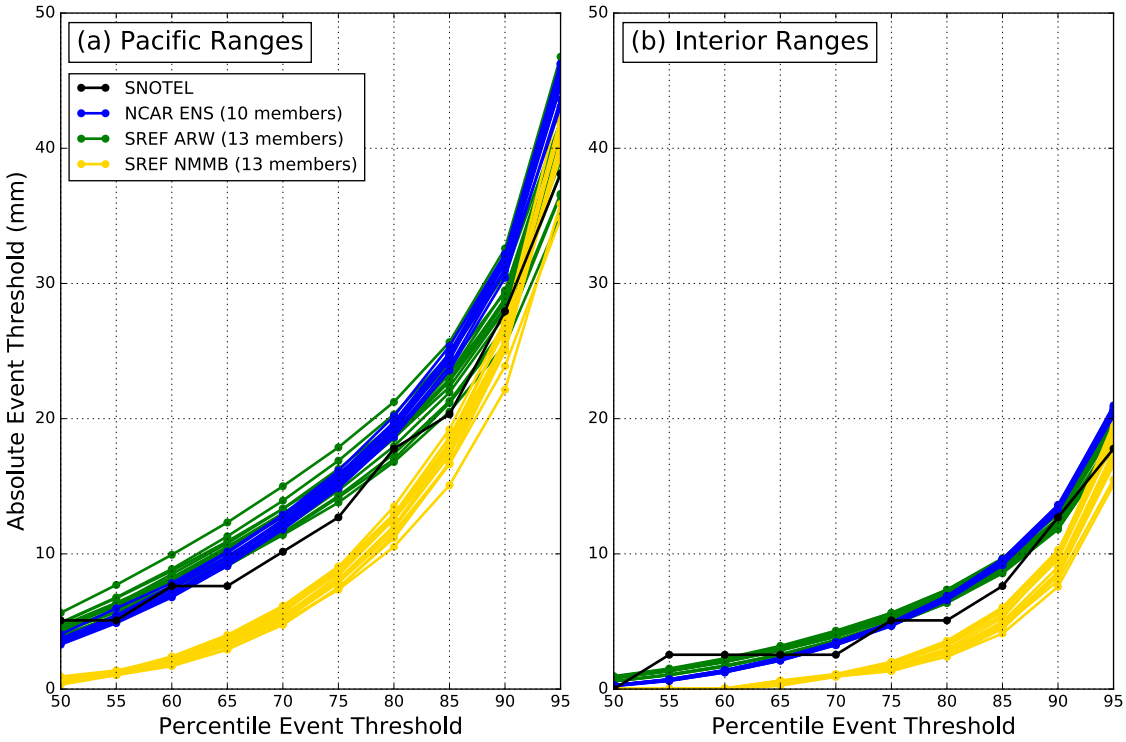
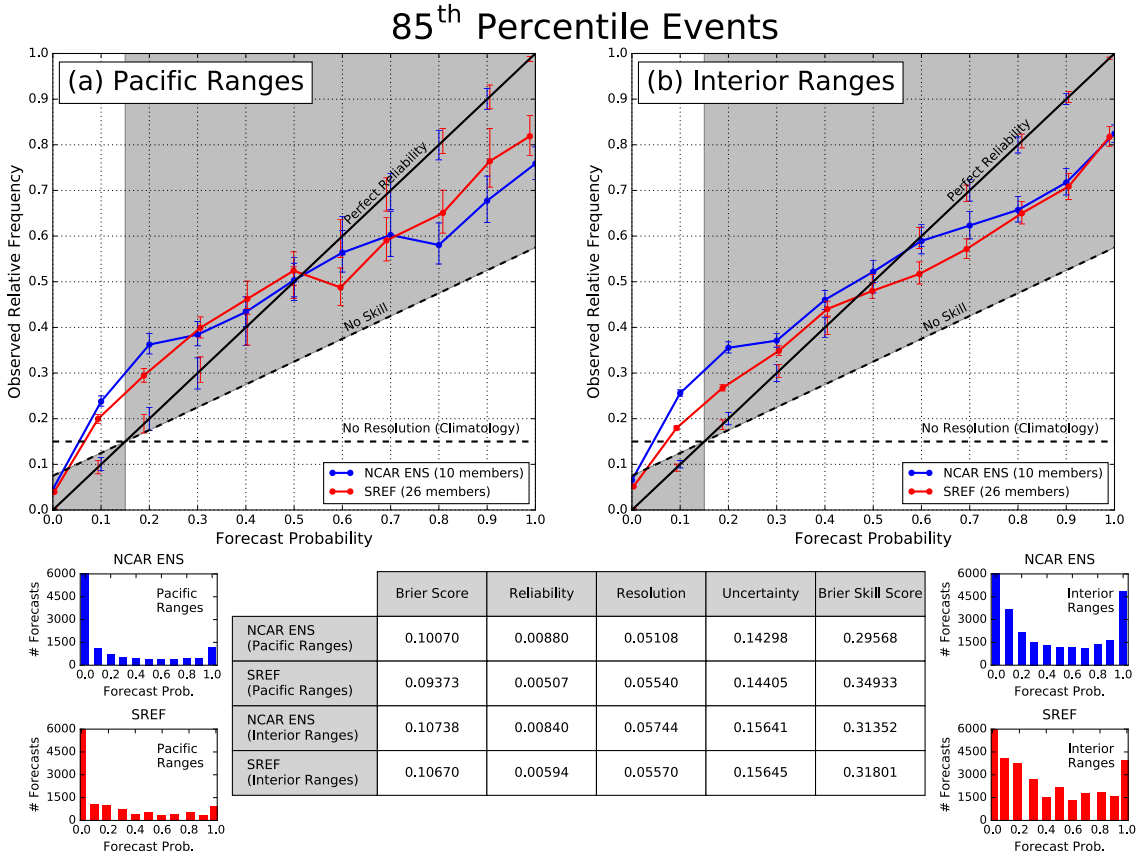
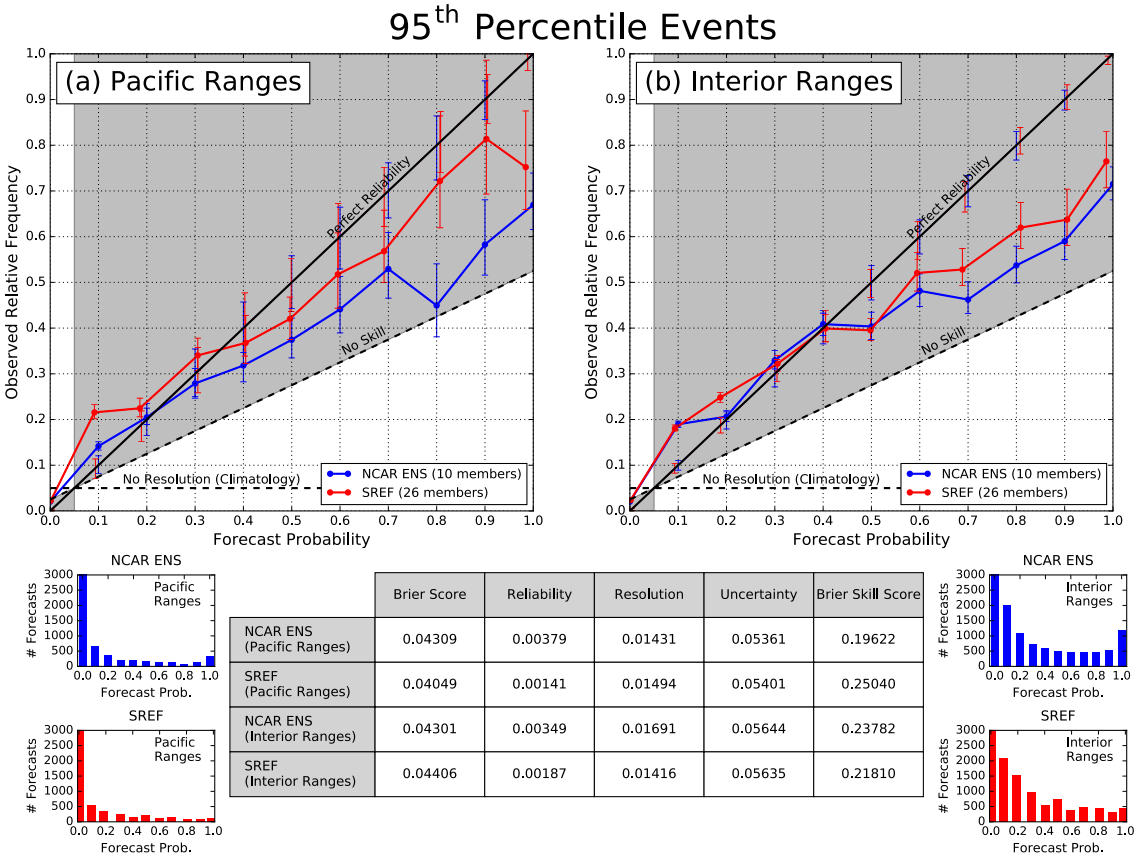


Figure 3.14. Same as Figure 3.12 except for all members of the NCAR ENS, SREF ARW, and SREF NMMB.



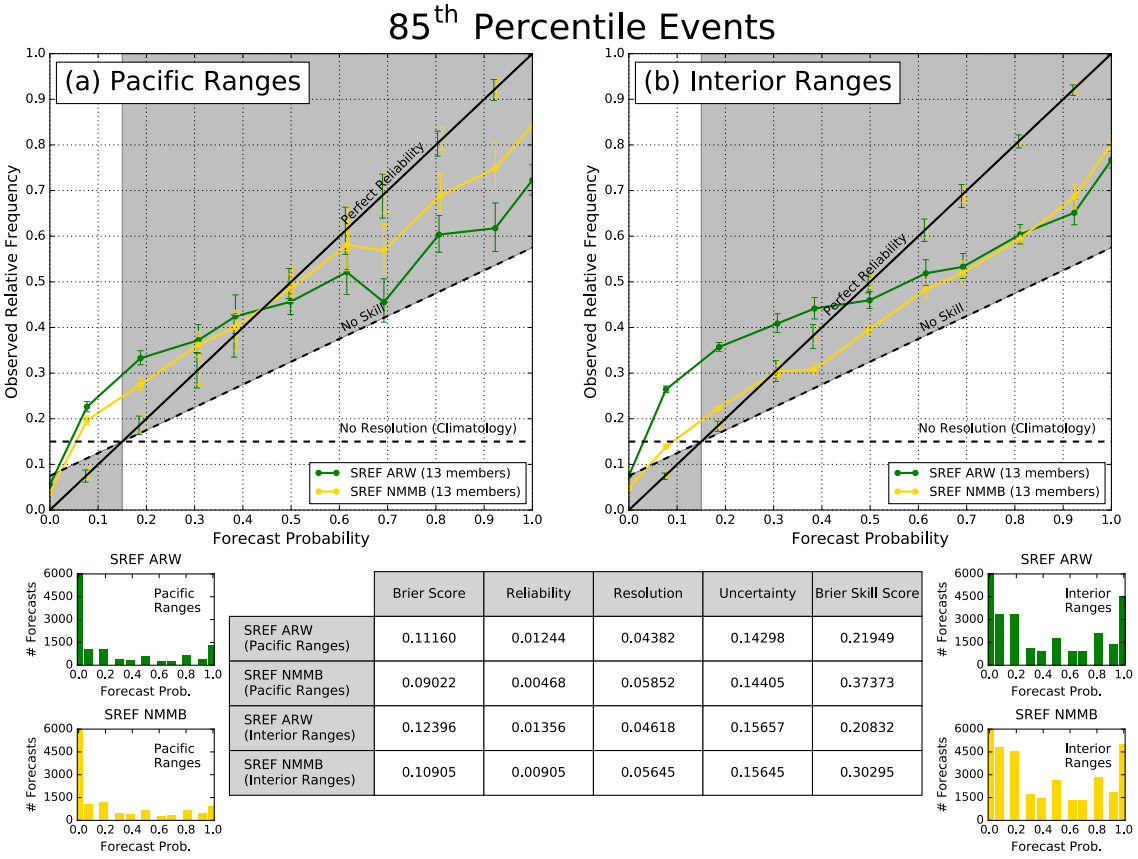
	Brier Score	Reliability	Resolution	Uncertainty	Brier Skill Score
NCAR ENS (Pacific Ranges)	0.10070	0.00880	0.05108	0.14298	0.29568
SREF (Pacific Ranges)	0.09373	0.00507	0.05540	0.14405	0.34933
NCAR ENS (Interior Ranges)	0.10738	0.00840	0.05744	0.15641	0.31352
SREF (Interior Ranges)	0.10670	0.00594	0.05570	0.15645	0.31801

Figure 3.15. Attributes diagram for NCAR ENS and SREF forecasted and SNOTEL observed 85th percentile events in the (a) Pacific ranges and (b) interior ranges. Histograms at bottom left (right) correspond to Pacific (interior) ranges and indicate number of forecasts in each forecast probability bin.



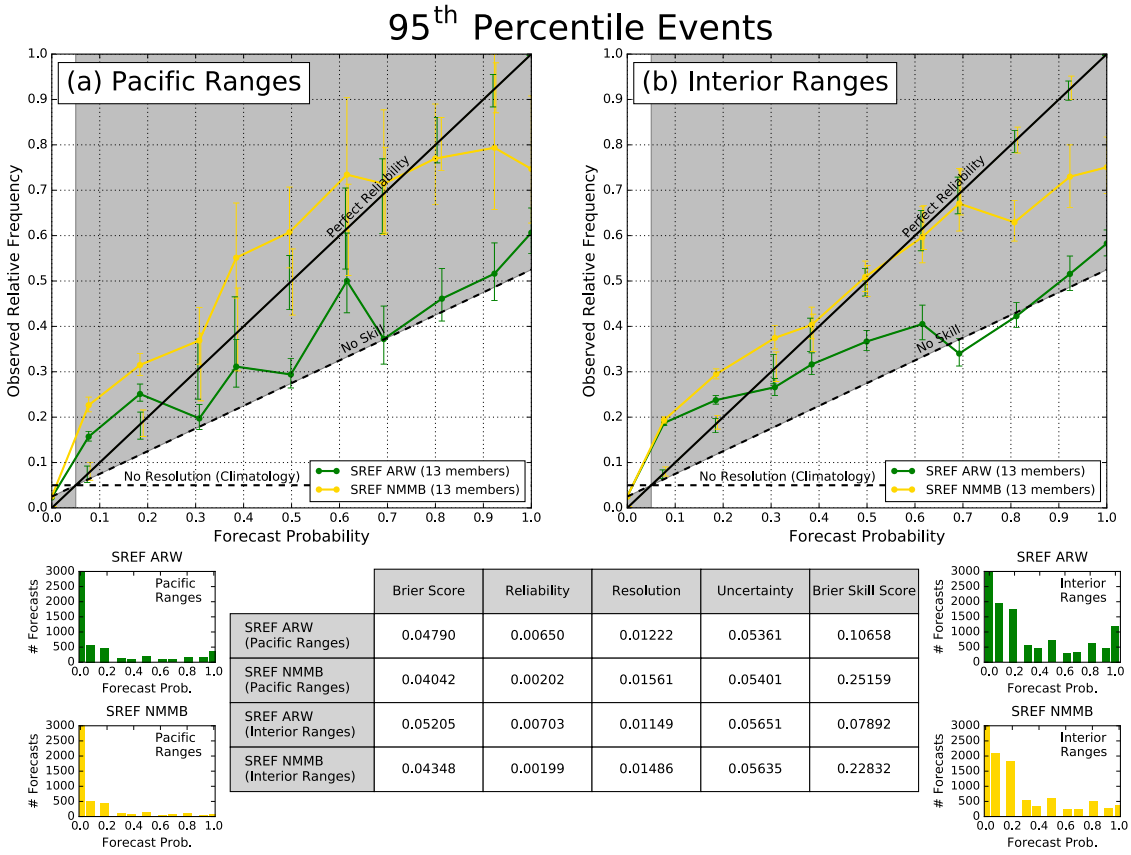
	Brier Score	Reliability	Resolution	Uncertainty	Brier Skill Score
NCAR ENS (Pacific Ranges)	0.04309	0.00379	0.01431	0.05361	0.19622
SREF (Pacific Ranges)	0.04049	0.00141	0.01494	0.05401	0.25040
NCAR ENS (Interior Ranges)	0.04301	0.00349	0.01691	0.05644	0.23782
SREF (Interior Ranges)	0.04406	0.00187	0.01416	0.05635	0.21810

Figure 3.16. Same as Figure 3.15 except for 95th percentile events.



	Brier Score	Reliability	Resolution	Uncertainty	Brier Skill Score
SREF ARW (Pacific Ranges)	0.11160	0.01244	0.04382	0.14298	0.21949
SREF NMMB (Pacific Ranges)	0.09022	0.00468	0.05852	0.14405	0.37373
SREF ARW (Interior Ranges)	0.12396	0.01356	0.04618	0.15657	0.20832
SREF NMMB (Interior Ranges)	0.10905	0.00905	0.05645	0.15645	0.30295

Figure 3.17. Same as Figure 3.15 except for SREF ARW and SREF NMMB.



	Brier Score	Reliability	Resolution	Uncertainty	Brier Skill Score
SREF ARW (Pacific Ranges)	0.04790	0.00650	0.01222	0.05361	0.10658
SREF NMMB (Pacific Ranges)	0.04042	0.00202	0.01561	0.05401	0.25159
SREF ARW (Interior Ranges)	0.05205	0.00703	0.01149	0.05651	0.07892
SREF NMMB (Interior Ranges)	0.04348	0.00199	0.01486	0.05635	0.22832

Figure 3.18. Same as Figure 3.15 except for SREF ARW and SREF NMMB and 95th percentile events.

CHAPTER 4

CONCLUSION

This study has evaluated the performance of precipitation forecasts from the convection-permitting NCAR ENS and several operational forecast systems at high-elevation SNOTEL sites across the western U.S. during the 2016/17 cool-season. The NCAR ENS CTL and HRRR exhibited superior precipitation biases as evinced by the ratio of forecast to observed mean daily precipitation and the ratio of forecast to observed event frequencies. Because it was effectively a combination of two short-term forecasts, the HRRR may have had an advantage. The GFS and SREF ARW CTL produced minimal overall bias, but overpredict or underpredict precipitation on a site by site basis. A significant wet bias is present in the NAM-3km due to its tendency to produce too many large events, especially over the interior ranges for events ≥ 20 mm, whereas the SREF NMMB CTL generates too few moderate and small events ≤ 20 mm over both regions, giving it a substantial dry bias.

Deterministic validation metrics (i.e., equitable threat scores, hit rates, and false alarm ratios) using absolute event thresholds indicate that the higher resolution NCAR ENS CTL, HRRR, and NAM-3km generally perform better than the coarser GFS, SREF ARW CTL, and SREF NMMB CTL. One exception is the performance of the NCAR ENS CTL over the Pacific ranges, where it exhibits poorer ETSs and false alarm ratios than the GFS.

This may reflect the close proximity of the NCAR ENS CTL's 3-km boundary to the west coast of the U.S (Schwartz et al. 2015). The SREF ARW CTL generally performs second worst for all 3 metrics, while SREF NMMB CTL produces the worst scores by a significant margin for all 3 metrics in both regions. Consistent with other studies (e.g., Lewis et al. 2016), the performance of all six models declines from the Pacific to interior ranges.

We further bias-correct these deterministic validation metrics by using percentile event thresholds. The removal of bias allows for a robust assessment of the spatial placement of precipitation within the context of each model's climatology. Overall, the bias-corrected results are generally consistent with the non-bias-corrected results when accounting for the impact that bias has on these three statistical measures. For example, although the bias-corrected ETSs are slightly lower for models with a wet bias (i.e., the NAM-3km), we still find the HRRR, NAM-3km, and GFS to exhibit the highest ETSs over the Pacific ranges and the HRRR, NAM-3km, and NCAR ENS CTL to exhibit the highest ETSs over the interior ranges.

Prior studies noted varied results concerning the benefits of decreasing grid spacing below 12 km over the western U.S. (Mass et al. 2002; Grubišić et al. 2005; Hart et al. 2005). Our results indicate that decreasing horizontal grid spacing to 3-km increases the performance of cool-season QPFs, especially over the interior ranges of the western U.S. The importance of increased resolution over the interior ranges may reflect their narrow nature, whereas the Pacific ranges have a more sustained high-mountain mass and are better resolved at coarser resolutions.

Although the NCAR ENS and SREF are both designed to produce short-range, probabilistic forecasts, their configurations, characteristics, and biases are drastically

different. While the NCAR ENS contains a single dynamical core and identical physics in each member, the SREF contains two dynamical cores (SREF ARW and SREF NMMB) with varied physics among the members in each core. Ideally, each member of an ensemble should be equally likely to be correct, and, thus, all members should have identical climatologies. We find the precipitation climatology for each member of the NCAR ENS to be similar, whereas the precipitation climatologies for the SREF bifurcate into two distinct clusters based on dynamical cores. While the NCAR ENS confirms the expectation of equal likelihood due to EAKF initializations, the design of the SREF clearly violates this principal. Consistent with the biases of their control members, NCAR ENS members contain a slight wet bias for 80th percentile and larger events, SREF ARW members contain an overall slight wet bias, and SREF NMMB members exhibit a significant dry bias, especially for 85th percentile events and smaller.

Bias-corrected probabilistic validation metrics reveal that although the NCAR ENS is generally more skillful than the SREF's individual dynamical cores, the full 26-member SREF commonly outperforms the NCAR ENS. Over the Pacific ranges, the NCAR ENS is less skillful than the SREF for both 85th and 95th percentile event thresholds. Meanwhile, over the interior ranges, the NCAR ENS exhibits more skill for 95th percentile event thresholds. The poorer relative performance of the NCAR ENS over the Pacific ranges is consistent with deterministic results. Overall, the NCAR ENS has slightly better resolution than the SREF, indicating that it is better at discriminating when an event occurs with lower or higher frequency than climatology. Probabilistic forecasts from the NCAR ENS are characterized by excessive sharpness, overconfidence, and poor reliability, whereas the SREF is less sharp and more reliable. Compared to individual SREF dynamical cores

(SREF ARW and SREF NMMB), the NCAR ENS is more skillful than the SREF ARW over the entire western US and the SREF NMMB over the interior ranges. By combining two ensemble systems with drastically different climatologies, the SREF is able to generate greater spread than the NCAR ENS, leading to probabilistic forecasts that are generally more skillful.

These findings indicate the advantages of high-resolution deterministic models and future promise of CPEs over the western U.S. The HRRR, NAM-3km, and NCAR ENS CTL consistently outperform the coarser GFS, SREF ARW CTL, and SREF NMMB, especially over the interior ranges. As computational resources increase, future work should focus on the development of operational deterministic models with horizontal grid spacings of 3 km or smaller. Although the NCAR ENS suffers from spread deficiency, its configuration should serve as a framework for the future development of short-range ensembles. With a horizontal grid spacing of 3 km, an individual member of the NCAR ENS is shown to be much more skillful than individual members of the 16-km SREF and, because it follows the principal of equal likelihood, its probabilistic forecasts can be easily interpreted. The NCAR ENS's downfall is insufficient spread, which hinders the performance of its probabilistic forecasts. Therefore, future work should specifically focus on improving spread in high-resolution, single-physics, single-dynamical core EPSs. As computational resources increase, a simple solution to this problem is to increase the number of ensemble forecast members.

REFERENCES

Alexander, C., and Coauthors, 2011: The High Resolution Rapid Refresh (HRRR): Recent and future enhancements, time-lagged ensembling, and 2010 forecast evaluation activities. *24th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction, Seattle, WA*. [Available online at https://ams.confex.com/ams/91Annual/webprogram/Paper183065.html.]

Alexander, C., and Coauthors, 2014: The High-Resolution Rapid Refresh (HRRR): A maturation of frequently updating convection-allowing numerical weather prediction. *Extended Abstracts, World Weather Open Science Conf., Montreal, QC, Canada, CIRES/ESRL/NCEP, 45 pp.* [Available online at https://www.wmo.int/pages/prog/arep/wwrp/new/wwosc/documents/WWOSC2014_Alexander_Final.pdf.]

Anderson, J.L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, doi:10.1175/1520-0493(2001)129,2884:AEAKFF.2.0.CO;2.

Anderson, J.L., 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634–642, doi:10.1175/ 1520-0493(2003)131,0634:ALLSFF.2.0.CO;2.

Barrett, A.I., S.L. Gray, D.J. Kirshbaum, N.M. Roberts, D.M. Schultz, and J.G. Fairman, 2016: The utility of convection-permitting ensembles for the prediction of stationary convective bands. *Mon. Wea. Rev.*, **144**, 1093–1114, doi:10.1175/MWR-D-15-0148.1

Ben Bouallègue, Z., S.E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorol. Z.*, **22**, 49–59, doi:10.1127/0941-2948/2013/0374.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, doi:10.1175/MWR-D-12-00031.1.

Brill, K.F., 2009: A general analytic method for assessing sensitivity to bias of performance measures for dichotomous forecasts. *Wea. Forecasting*, **24**, 307–318, doi:10.1175/2008WAF2222144.1.

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev*., **78**, 1–3, doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brocker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**,651–661, doi:10.1175/ WAF993.1.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface-hydrology model with the Penn State–NCARMM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585, doi:10.1175/1520-0493(2001)129,0569: CAALSH.2.0.CO;2.

Clark, A.J., W.A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, doi:10.1175/2009WAF2222222.1.

Clark, A.J., and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, doi:10.1175/2010MWR3624.1.

Clark, P., Roberts, N., Lean, H., Ballard, S.P. and Charlton-Perez, C. 2016: Convection-permitting models: A step-change in rainfall forecasting. *Met. Apps.*, **23**, 165–181. doi:10.1002/met.1538.

Colle, B.A., 2004: Sensitivity of orographic precipitation to changing ambient conditions and terrain geometries: An idealized modeling perspective. *J. Atmos. Sci.*, **61**, 588–606, doi:10.1175/1520-0469(2004)061<0588:SOOPTC>2.0.CO;2.

Colle, B.A., M.F. Garvert, J.B. Wolfe, C.F. Mass, and C.P. Woods, 2005: The 13–14 December 2001 IMPROVE-2 event. Part III: Simulated microphysical budgets and sensitivity studies. *J. Atmos. Sci.,* **62**, 3535–3558, doi:10.1175/JAS3552.1.

Daly, C., R.P. Neilson, and D. L. Phillips, 1994: A statistical–topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158, doi:10.1175/15200450(1994)033<0140:ASTMFM>2.0.CO;2.

Daly, C., M. Halbleib, J.I.Smith,W.P.Gibson, M. K. Doggett, G. H. Taylor,J.Curtis, andP.P.Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, doi:10.1002/joc.1688.

Dey, S.R., G. Leoncini, N.M. Roberts, R.S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, doi:10.1175/MWR-D-14-00172.1

Di Luzio, M., G.L. Johnson, C. Daly, J.K. Eischeid, and J.G. Arnold, 2008: Constructing

retrospective gridded daily precipitation and temperature datasets for the conterminous united states. *J. Appl. Meteor. Climatol.*, **47**, 475–497, doi:10.1175/2007JAMC1356.1.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Short Range Ensemble Forecast (SREF) system at NCEP: Recent development and future transition. *23rd Conf. on Numerical Weather Prediction/27th Conf. on Weather Analysis and Forecasting*, Chicago, IL, Amer. Meteor. Soc., 2A.5. [Available online at https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html.]

Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, doi:10.1002/met.25.

Efron, B., and R.J. Tibshirani, 1993: An introduction to the bootstrap. Chapman & Hall/CRC, 456 pp.

Fassnacht, S.R., 2004: Estimating Alter-shielded gauge snowfall undercatch, snowpack sublimation, and blowing snow transport at six sites in the coterminous USA. *Hydrol. Processes*, **18**, 3481–3492, doi:10.1002/hyp.5806.

Gallo, B.T., A.J. Clark, and S.R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, doi:10.1175/WAF-D-15-0134.1.

Garvert, M.F., C.P. Woods, B.A. Colle, C.F. Mass, P.V. Hobbs, M.T. Stoelinga, and J.B. Wolfe, 2005: The 13–14 December 2001 IMPROVE-2 Event. Part II: Comparisons of MM5 model simulations of clouds and precipitation with observations. *J. Atmos. Sci.*, **62**, 3520–3534, doi:10.1175/JAS3551.1.

Gebhardt, C., Theis, S.E., Paulat, M., Ben Bouallègue, Z., 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, doi:10.1016/j.atmosres.2010.12.008.

Grubišić, V., R.K. Vellore, and A.W. Huggins, 2005: Quantitative precipitation forecasting of wintertime storms in the Sierra Nevada: Sensitivity to the microphysical parameterization and horizontal resolution. *Mon. Wea. Rev.*, **133**, 2834–2859, doi:10.1175/MWR3004.1.

Hamill, T.M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Hart, K.A., W.J. Steenburgh, and D.J. Onton, 2005: Model forecast improvements with decreased horizontal grid spacing over finescale intermountain orography during

the 2002 Olympic Winter Games. *Wea. Forecasting*, **20**, 558–576, doi:10.1175/WAF865.1.

Iacono, M.J., J.S. Delamere, E.J. Mlawer, M.W. Shephard, S.A. Clough, and W.D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, doi:10.1029/ 2008JD009944.

Ikeda, K., and Coauthors, 2010: Simulation of seasonal snowfall over Colorado. *Atmos. Res.*, **97**, 462–477, doi:10.1016/j.atmosres.2010.04.010.

Jirak, I.L., S.J. Weiss, and C.J. Melick, 2012: The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *Preprints, 26th Conf. Severe Local Storms*, Nashville, TN. [Available online at http://www.spc.noaa.gov/publications/jirak/disagg.pdf.]

Jirak, I.L., C.J. Melick, and S.J. Weiss, 2016: Comparison of the SPC Storm-Scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. *Preprints, 28th Conf. Severe Local Storms, Portland, OR*. [Available online at http://www.spc.noaa.gov/publications/jirak/sseocomp.pdf.]

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, doi:10.1175/1520-0493(1994)122,0927:TSMECM.2.0.CO;2.

Janić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in thence Meso model. NCEP Office Note 437, 61 pp. [Available online at http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.]

Johnson, A. and X. Wang, 2016: A study of multiscale initial condition perturbation methods for convection-permitting ensemble forecasts. *Mon. Wea. Rev.,* **144**, 2579–2604, doi:10.1175/MWR-D-16-0056.1.

Kain, J.S., S.J. Weiss, D.R. Bright, M.E. Baldwin, J.J. Levit, G.W. Carbin, C.S. Schwartz, M.L. Weisman, K.K. Droegemeier, D.B. Weber, and K.W. Thomas, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. Wea. Forecasting, 23, 931–952, doi.org/10.1175/WAF2007106.1.

Mass, C.F., D. Ovens, K. Westrick, and B.A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts*?*. *Bull. Amer. Meteor. Soc.*, **83**, 407–430, doi.:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Melhauser, C., F. Zhang, Y. Weng, Y. Jin, H. Jin, and Q. Zhao, 2017: A multiple-model convection-permitting ensemble examination of the probabilistic prediction of

tropical cyclones: Hurricanes Sandy (2012) and Edouard (2014). *Wea. Forecasting*, **32**, 665–688, doi:10.1175/WAF-D-16-0082.1.

Murphy, A.H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, doi.:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504, doi:10.1175/WAF-D-13-00066.1.

Le Duc, L., K. Saito, and H. Seko, 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts, *Tellus A*, **65**, 18171, doi:10.3402/tellusa.v65i0.1817.

Lewis, W.R., W.J. Steenburgh, T.I. Alcott, and J.J. Rutz, 2017: GEFS precipitation forecasts and the implications of statistical downscaling over the western united states. *Wea. Forecasting*, **32**, 1007–1028, doi:10.1175/WAF-D-16-0179.1.

Lorenz, E.N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, doi:10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.*, **37**, 75–81.

Mason, I., 2003: Binary events. *Verification: A Practitioner's Guide in Atmospheric Science*, I.T. Jolliffe and D.B. Stephenson, Eds., John Wiley and Sons, 37–76.

Mellor, G.L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys. Space Phys.,* **20**, 851–875, doi:10.1029/RG020i004p00851.

Mlawer, E. J., S.J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the long-wave. *J. Geophys. Res.*, **102**, 16 663–16 682, doi:10.1029/97JD00237.

Munsell, E.B., J.A. Sippel, S.A. Braun, Y. Weng, and F. Zhang, 2015: Dynamics and predictability of hurricane nadine (2012) evaluated through convection-permitting ensemble analysis and forecasts. *Mon. Wea. Rev.*, **143**, 4514–4532, doi:10.1175/MWR-D-14-00358.1.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Rasmussen, R., and Coauthors, 2012: How well are we measuring snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed. Bull. Amer. Meteor. Soc., 93, 811–829, doi:10.1175/ BAMS-D-11-00052.1.

Roberts, N.M. and H.W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Roe, G.H., 2005: Orographic precipitation. *Ann. Rev. Earth Planet. Sci.*, **33(1)**, 645–671, doi:10.1146/annurev.earth.33.092203.122541.

Rogers, E., and Coauthors, 2017: Mesoscale modeling development at the National Centers for Environmental Prediction: Version 4 of the NAM Forecast System and scenarios for the evolution to a high-resolution ensemble forecast system. *28th Conf. on Weather and Forecasting/24th Conf. on Numerical Weather Prediction*, Seattle, WA. [Available online at https://ams.confex.com/ams/97Annual/webprogram/Paper311212.html.]

Romine, G.S., C.S. Schwartz, J. Berner, K.R. Fossell, C. Snyder, J.L. Anderson, and M.L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, doi:10.1175/MWR-D-14-00100.1.

Rotunno, R. and Houze, R.A., 2007: Lessons on orographic precipitation from the Mesoscale Alpine Programme. *Q.J.R. Meteorol. Soc.*, **133**: 811–830. doi:10.1002/qj.67.

Rutz, J.J., and W. J. Steenburgh, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905–921, doi:10.1175/MWR-D-13-00168.1.

Rutz, J.J., W.J. Steenburgh, and F.M. Ralph, 2015: The inland penetration of atmospheric rivers over western North America: A Lagrangian analysis. *Mon. Wea. Rev.*, **143**, 1924–1944, doi:10.1175/MWR-D-14-00288.1

Schaefer, J.T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, doi:10.1175/1520-0434(1990)005,0570:TCSIAA.2.0.CO;2.

Schellander-Gorgas, T., Wang, Y., Meier, F., Weidle, F., Wittmann, C., and Kann, A., 2017: On the forecast skill of a convection permitting ensemble. *Geosci. Model Dev.*, **10**, 35–56, doi:10.5194/gmd-10-35-2017.

Schwartz, C.S., and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**,

3351–3372, doi:10.1175/2009MWR2924.1.

Schwartz, C.S., 2014: Reproducing the September 2013 record-breaking rainfall over the Colorado Front Range with high-resolution WRF forecasts. *Wea. Forecasting*, **29**, 393–402, doi:10.1175/WAF-D-13-00136.1.

Schwartz, C.S., G.S. Romine, R.A. Sobash, K.R. Fossell, and M.L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi:10.1175/WAF-D-15-0103.1.

Serreze, M.C., M.P. Clark, R.L. Armstrong, D. A. McGinnis, and R. S. Pulwarty, 1999: Characteristics of the western United States snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **35**, 2145–2160, doi:10.1029/1999WR900090.

Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob, and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *J. Geophys. Res.*, **102**, 23 895–23 915, doi:10.1029/97JD01864.

Thompson, G., P.R. Field, R.M. Rasmussen, and W.D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, doi:10.1175/2008MWR2387.1.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1779–1800, doi:10.1175/1520-0493(1989)117,1779:ACMFSF.2.0.CO;2.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. Verification: A Practitioner's Guide in Atmospheric Science, I.T. Jolliffe and D.B. Stephenson, Eds., John Wiley and Sons, 137–163.

Trier, S.B., G.S. Romine, D.A. Ahijevych, R.J. Trapp, R.S. Schumacher, M.C. Coniglio, and D.J. Stensrud, 2015: Mesoscale thermodynamic influences on convection initiation near a surface dryline in a convection-permitting ensemble. *Mon. Wea. Rev.*, **143**, 3726–3753, doi:10.1175/MWR-D-15-0133.1

Vié, B., Molinié, G., Nuissier, O., Vincendon, B., Ducrocq, V., Bouttier, F., and Richard, E., 2012: Hydro-meteorological evaluation of a convection-permitting ensemble prediction system for Mediterranean heavy precipitating events, *Nat. Hazards Earth Syst. Sci.*, **12**, 12, 2631–2645, doi:10.5194/nhess-12-2631-2012.

Weisman, M.L., C.A. Davis, W. Wang, K.W. Manning, and J.B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, doi:10.1175/2007WAF2007005.1.

Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences. 3rd ed. International Geophysics Series*, Vol. 100, Academic Press, 676 pp.

Yang, D., B.E. Goodison, J.R. Metcalfe, V. S. Golubev, R. Bates, T. Pangburn, and C. L. Hanson, 1998: Accuracy of NWS 8" standard nonrecording precipitation gauge: Results and application of WMO intercomparison. *J. Atmos. Oceanic Technol.*, **15**, 54–68, doi::10.1175/1520-0426(1998)015<0054:AONSNP>2.0.CO;2.

Zhang, F. and Y. Weng, 2015: Predicting hurricane intensity and associated hazards: A five-year real-time forecast experiment with assimilation of airborne doppler radar observations. *Bull. Amer. Meteor. Soc.*, **96**, 25–33, doi:10.1175/BAMS-D-13-00231.1.